Humanities, Computers and Cultural Heritage

Proceedings

of the XVI international conference
of the Association for History and Computing

14-17 September 2005

Royal Netherlands Academy of Arts and Sciences
Amsterdam, 2005

ISBN 90-6984-456-7

The paper in this publication meets the requirements of ∞ iso-norm 9706 (1994) for permanence.

Design
Edita-KNAW, Amsterdam
www.knaw.nl/edita

Communication
www.ahc2005.org

Low Countries Organisation Committee
Onno Boonstra, Humanities computing, University of Nijmegen
Leen Breure, Computer and Information Science, University of Utrecht
Peter Doorn, Data Archiving and Networked Services (DANS), The Hague
Jaap van den Herik, Computer Science, Universities of Leiden and Limburg
Bart de Nil, Institute for Social History (Amsab), Gent, Belgium
Paula Witkamp, European Commission on Preservation and Access, Amsterdam

Local organisation
Michelle van den Berk, Berry Feith, Yamit Gutman and Annelies van Nispen
(Netherlands Institute for Scientific Information Services – NIWI-KNAW)

Organizing institutions
AHC: international Association for History and Computing
KNAW: Royal Netherlands Academy of Arts and Sciences
NIWI: Netherlands Institute for Scientific Information Sciences
VGI: Low Countries branch of the Association for History and Computing
SIKS: Dutch research school for Information and Knowledge Systems
DANS: Data Archiving and Networked Services

Preface

'History & Computing' and 'Humanities Computing' are in a crucial phase of development. The cultural heritage sector is turning digital, and ever more archival and other historically relevant sources are becoming available online. As a result there is a need for innovative methods and techniques to process the flood of digital resources. Currently, computer scientists show a keen interest in the information problems of cultural heritage and the humanities. Advances in grid computing and the Semantic Web are stimulating a new kind of e-Science. e-Culture, e-Humanities and even e-History.

This volume contains about fifty contributions presented at the XVIth international conference of the Association for History and Computing, that took place in Amsterdam from September 14th till 17th 2005.

The conference papers are intended for an audience of specialists from three broad fields:

- Scholars using computers in historical and related studies (history of art, archaeology, literary studies, etc.)
- Information and computing scientists working in the domain of cultural heritage and the humanities
- Professionals working in cultural heritage institutes (archives, libraries, museums) who use ICT to preserve and give access to their collections

The subject matter of these proceedings is primarily oriented at methodological issues. It is not restricted to one particular domain within history and the humanities. The papers included in this volume were selected by the conference committee and the session convenors. Due to time pressure, the papers could only superficially be refereed and marginally edited. Some papers that arrived too late to be processed could regrettably not be included in the volume. Such papers and the abstracts of poster presentations can be found on the conference website (www.ahc2005.org). A selection of the proceedings is being considered for publication in international journals.

The papers in this volume of proceedings can be characterized on the following characteristics:

- A first group of papers deals with *portals* and *gateways* to heritage information. More particularly, several papers on *virtual libraries* and *digital archives* are included.
- *Data enrichment* is the overall theme of the papers on on *electronic text editing* and digital source editions. *Text analysis and retrieval* is a subject that has always received some, but not much attention from computing historians. We are happy that in this volume, a number of papers is dedicated to text analytical and ontological problems, partly inspired by the discussions

on the Semantic Web, but also influenced by participants from the field of literary and linguistic computing.
- *Images & multimedia* is one of the subject that attracts special attention from computing scientists. Papers on visual object detection and content-based artist identification show advances that are made in this area.
- *Geographical Information Systems* is a topic that has become remarkably 'hot' in historical studies over the past few years. The conference includes five sessions and 14 papers on historical GIS applications, ranging from building a historical GIS to time-space analysis and applications in urban history.
- *Quantitative data analysis* was one of the most important subjects in the early years of historical computing, but now it is attracting a relatively modest attention. Computer applications of statistics have become mainstream in social and economic history and apparently require less specific attention at AHC conferences.
- A number of papers deals with *digitization strategies* in heritage institutions and on the digitization of historical sources. However, the session convenors and other referees were fairly strict on referring 'me and my database' kind of papers to poster sessions, unless they clearly presented new methods of database design. The *XML markup language* offers a strong tool for the encoding of irregular source structures. A score of papers is dedicated to the role of XML in the structuring of heritage information.
- Large cross-sectional, nominative *databases in historical research* is a subject that might be called 'traditional' at AHC conferences. Two sessions and a handful of papers were dedicated to this subject.
- Finally, there is a number of papers with a *theoretical and methodological* component, in which *virtual networks* and collaboratories play a role. Moreover, several papers claiming *new approaches* to history and computing are included in these proceedings.

Peter Doorn
President of the AHC

Contents

Digitizing cultural heritage collections: The importance of training

Patricia Alkhoven

Many Cultural Heritage Institutions are starting to digitize parts of their collections. They have often very few staff available for these 'new' activities. Often without any training or preparation they start scanning objects. Very little attention is payed to standards, interoperability, workflow efficiency, cooperation with other institutes or even the quality of the product in terms of scholarly content, image quality, usability etc. This paper explanes why training is necessary before huge investments in time and money are made and end-products appear to be disappointing.

Introduction

Many cultural heritage institutions are starting to digitize parts of their collections for better accessibility, preservation of originals, promotion and visibility of the institution on the Internet, competition with other institutions, etc. Watching other cultural heritage institutions showing their collections on the Internet, the pressure is high to do the same. They have often very few staff available for these 'new' activities. A director decides that his institution should present itself as well on the Internet and give direct access to (part of) their collections. The documentalist, keeper of special collections or librarian is asked to start scanning objects, often without any training or preparation. The focuss is very much on the scanning, while scanning is only part of the whole digitization process. Often not enough attention is being given tot de workflow and the analysis of the materials to be digitized. As a result, very little attention is payed to standards, interoperability, workflow efficiency, cooperation with other institutions or even the quality of the product in terms of scholarly content, image quality, usability etc.

Fact is that most digitization projects largely overspend their budgets, are launched later than expected, require more staff and much more attention is being needed to keep and update the data. So, why is this? One would say that if a sound estimate of the expected costs based on realistic figures has been made, the project would likely stay within budget and meet the deadlines. Reality shows differently.

As we inventory the results of recent digitization projects, it is clear that there is a wide spectrum of websites of cultural heritage institutions. They differ in every sense: target groups, metadata, appearance, navigation, content etc. There is a lot of fragmentation and loss of efforts since institutions work together only in very few cases. Exchangeability and interoperability have also not the highest priority as appears from the many different ways metadata have been used, different soft- and hardware programs, standards, etc.

It is a fact that digitization costs a lot, always. Smaller cultural heritage institutions can hardly afford the extra expenses for digitization. They have few stafmembers available that can spend time on a digitization project and they have little means for maintenance or keeping the information up-to date.

For scanning, volonteers are often hired to do the job. In most of the cases they have no training or expertise to speak of in that area. Although the scans from such projects can be acceptible in quality, without proper projectmanagement the project may fail al-

together and at least takes longer and costs more.

Underestimated are:

- Costs per scan versus total costs
- Linking image + description
- Analysis of materials
- Projectmanagement
- The need for quality control
- Digital preservation
- Maintenance / updates

Costs per scan versus total costs

It would not be the first time that a cultural heritage director says: 'I know a company where we pay 10 Eurocent per scan, so if we digitize 100.000 pages it will cost me about 10.000 Euro'! This sounds very naïve, because the price mentioned only includes the raw scan and not the handling before or after scanning. No filemanagement, the creation of derivatives (from TIFF to JPEG), image manipulation such as cropping, cutting, enhancing or storage on CD's or any other medium is included. The total scanning costs make up about 30% of the total project costs. This means that finally a scan will cost many times more than the 10 eurocents per scan.

Linking image + description

The linking of images and descriptions (records) is often underestimated since apparently many people still think that once they have an image and record that's all they need, and 'the computer will do the rest'! That the record and the images need to be linked by an number, filename or url is basically not much different from giving each book a shelfnumber and store the book accordingly in the storage facility. The linking is especially a human task that needs carefull attention and is therefore relatively expensive. In fact it is one of the most important task of the project: if something goes wrong with the linking it will be immediately visible on the web!

Analysis of materials

Just counting the number of volumes or mesuring the meters and estimating the amount of pages is not enough. Materials to be digitized can vary in size, completeness, they may be soiled or damaged and need conservation first etc. All cost are based on the total number of images. When the number of pages deviates too much from the estimated figure, it can have serious impact on the project. The size is also important: If there are different kinds of size it may be difficult and costly to have the scans made: the scanner has to adjust the camera for each individual shot. Sorting by size, type and material can help this process. In fact it is best to leaf through the books and check very carefully what kind of material it is and how many pages it has!

Projectmanagement

Projectmanagement is very important in a digitization project. Many parties are involved that need to be involved and 'speak' to each other. The projectmanager is often the translater between the experts requests and the technical possibilities, user requirements and institutional policies and external parties. Even with a project plan and detailed working plan the human factor of a projectmanager is crucial for the success or failure of a project.

The need for quality control

There is a significant difference between digitization only for simple access and digitization for preservation of cultural heritage. Digitization for simple access does not require the highest resolution of scans. They serve as a surrogates for the originals, that are still available and in good condition. Digitization for preservation on the other hand requires images with the highest standards, an efficient method of storage and retrieval. Digitization for preservation presupposes that the digital image will replace the original as well as possible, that the master copy is loslessly stored and kept for the future. Digital derivatives can be created from the masters for access purposes. Controll of the images is very important. Even if qualified companies carry out the scanning work, independent quality controll is needed.

Digital preservation / maintenance / updates

For some institutions the project ends when the web site with the digitized material is available on the web. Even if they do not maintain the data themselves or have other institutions taking care of that, updating the materials is certainly necessary from time to time since hardware platforms, software versions and plugins programmes supporting the websites are in a continuously changing process. The costs and activities needed for this need to be anticipated in the project plans.

Quality of project plans

The Netherlands' National Digitization program *the Memory of the Netherlands* (www.geheugenvannederland.nl/en) has published clear guidelines for project proposals to conform to the Memory. The Memory of The Netherlands, coordinated by KB, presents 45 (at this moment) collections from different cultural heritage institutions. The lists in the guidelines only need to be filled in by the institutions. In fact the complete structure for a project plan is in place, helping the filing institution. There is a back draw though. Since they have not written the project proposal all by themselves, but just filled in the empty spaces, they appear to be less involved and less aware of the steps to take in the workflow.

Funding Institutions such as Mondriaan Foundation have each year a certain amount of money avaialable for relatively small digitization projects. The reviewers of the project proposals have been complaining for years now about the low quality of project plans.

Academic Institutions can file digitization project proposals to the Netherlands Organization for Scientific Research (NWO). One of the main requirements for election is that the research proposed could not be done without the digitized resources. A scientific committee evaluates the proposals and often conclude that the scientific or scholarly part has been well described but that there is much uncertain in the digitization part. Many problems such as quantities, qualities, technical applications or copyrights are underestimated.

In general, it appears to be difficult to write good project proposals in line with each of the organizations and complete with figures and details. Over the years these organizations have become more critical and better acquainted with the details of digitization. Their requirements are now more strict especially with respect to the technical feasibility, the scholarly 'extra' value and sustainability of digital data over a longer period of time.

Our conclusion was that with more and better practical training in digitization by having them writing complete project plans we could teach them to be better prepared for real projects.

Training in digitization

For almost 10 years now, courses in scanning and digitization have been organized by Cornell (www.library.cornell.edu/), School for Scanning www.nedcc.org/); Digitisation Summer School in Glasgow (*www.hatii.arts.gla.ac.uk*/); Digital Futures Academy http://www.kcl.ac.uk/kdcs/digifutures.htm; and many other organizations. Recently more specializes courses have been organized in *e.g.* digital preservation, metadata, open access etc.

In The Netherlands, for some years the International Ticer School (known for its former International Summer School on the Digital Library) in Tilburg provide courses at Tilburg University Library. Most of the above mentioned courses focus on policies and management aspects of large scale digitization projects. The information given is often very theoretical and participants will not get acquainted with practical exercises or workshops. Although Cornell on the other hand offers the well-known practical scanning course for professional scanners, participants learn relatively little about the rest of the process. The Cornell course resulted in the well known handbook on digitization: Moving Theory into Practice (Anne Kenney and Oya Rieger). A very complete online tutorial accompanies the book. But, reading an online tutorial is still different from being tought in a classroom.

Training for cultural heritage institutions

In The Netherlands, several stafmembers of the larger cultural heritage institutions and university libraries followed the courses given at Cornell University in the United States. For most smaller institutions this was too expensive or too far away. Summer courses in digitization closer by were mostly aiming at the higher management of larger institutions. There was a need from many of the smaller organizations for a different kind of education in digitization. A more practical course, in which also the larger cultural and policy framework would be sketched, would be a welcome addition to the existing courses.

The first plans for setting up a structured, practical course that encompassed the whole digitization process started during an international conference in Utrecht 'Digitisation of European Cultural Heritage: Products, Principles and Techniques (21-23 October 1999) Utrecht University. Members from Koninklijke Bibliotheek, Digital Heritage Netherlands (DEN), Utrecht University Library, the European Commission on Preservation and Access, and Library of the University of Amsterdam decided it was time to set up a course for the smaller cultural heritage institutions inThe Netherlands. All the members are personally involved in digi-

tization projects and were willing to teach. However, none of the institutions could issue official diploma's or certificates. Therefore, GO-opleidingen, a non-profit organization for educating employees of libraries, archives and museums etc. was contacted, since their target group would be the same.

Since January 2004 GO-opleidingen organises twice a year a short course 'Digitization of Cultural Collections' in The Hague. (http://www.stgo.nl/kort/kort-dicc.htm). The course is given in cooperation with the Koninklijke Bibliotheek, Digital Heritage Nederland (DEN), Utrecht University Library, the European Commission on Preservation and Access, and Library of the University of Amsterdam.

The participants – 26 received a certificate up to today – evaluate the course very positively (8.5!). They state they have learned very much of the experiences of the teachers.

One of the advantages is that the teachers are directly involved in digitization projects and as a result they can give practical tips and talk about their experiences. Although part of the course discusses theory, participants are required to do many exercises and write a real project proposal with two or three others which they have to defend in class.

Most participants are more or less involved or starting digitization projects in their institutions. The course provides for them an opportunity to discuss their problems with others and with the teachers.

Contents of the course:
– Projectmanagement
– Selection
– Access
– Types of materials
– Digital preservation
– Maintenance and exploitation
– Scanning and quality control
– Budgets, planning and workflow
– Presentations of project plans

Result: the participants should be able to write a complete project proposal and be able to argument his choices. In fact, the course stimulated the cooperation between the university and (non-)governmental organizations.

Training at university level: Masters Book & Digital Media at Leiden University

During the nineties of the last century the university of Leiden had a course for unemployed historians: historical information processing. This course appeared a fruitfull education, since many people who followed this course found interesting jobs in cultural herige institutions and else. Many students started as trainees, proved themselves well and were offered jobs of a more permanent nature. Unfortunately, due to shifting of the budgets the couse was stopped in 1999. KB has educated at least 20 trainees in the years before. At least seven of them found a temporary or permanent job in KB.

September 2004 the new Mastercourse Book and Digital Media: Book & Byte started at the University of Leiden. The National Library of the Netherlands and Library of Leiden University started a contract to cooperate with Leiden University. The Master's Course includes: 'Digital Access to Cultural Heritage' which is being given by KB and the Foundation Digital Cultural Heritage (DEN).

This course deals with the construction of an infrastructure for making available in a digital form cultural heritage collections and knowledge domains. Working within the knowledge transmission cycle framework (production, distribution and consumption), emphasis is placed on the interaction between providers and users of digital information. This course aims to provide a wide perspective on the digital cultural landscape and includes some practical exercises as well.

This course derives much from the course given by GO as mentioned above, but is more theoretical in nature, although all aspecs in digitization are being discussed. It aims to challenge students to look beyond the existing practice and develop new solutions, projects and possibilities for accessing, sharing and participating in cultural heritage within the framework of new media applications.

Contents of course: Digital Access to Cultural Heritage:

– *Digital Cultural Heritage Landscape* Contents: Cultural Heritage Institutions, National and international management framework; Projects, online journals, discussionlists, handbooks, expertise centra.....

– *Search & Retrieval of Cultural Heritage Resources,* Contents: Projects & webexamples; user perspectives & expectations
– Emerging Technologies for Accessing Digital Information**,** Contents: Semantic Web, Thesauri, Taxonomies, Ontologies, Persistent and Perpetual Access
– *Large Scale Implementations: Two 'Memory' Cases,* Contents: American Memory and Memory of the Netherlands Projects. Legal Issues
– *Presentation & Interfacing & Usability***;** Contents: presentation; interfacing, usability, project management
– *Projects:* Materials, Project proposals, Project Planning, Costs, documentation.
– *The Future Digital Heritage Space ;* Contents: ICT influence on Organisations; Roadmap for Heritage Institutions
– *Excursion: Hands-on Experience: the Digital Production Line;* Contents: Scanning and Quality Control

Concluding remarks

Now that there is nothing 'new'anymore about digitization of collections, because everybody is doing it, there is a tendency to take too much for granted and underestimate the complexity of digitization projects.

Better training will not guarantee the success of a project, but it will certainly stimulate that organizations are well prepared for the whole process. It is better to invest in training and do it right from the start than end up with the frustration of an overspent budget, deadline not met and disappointing results.

Digitization always costs a lot of money. Be sure it is well spent!

Using a GIS to reconstruct the nineteenth century Lisbon parishes

Daniel R. Alves

Introduction

The changes that took place in the internal organisation of the territory in the last two centuries make the knowledge of the national administrative division and the understanding of its evolution essential in Portuguese Contemporary History, especially for comparing historical data over time.

This perspective is still clearer regarding the history of Lisbon. The demographic evolution of the last two centuries and the resulting growth of the city were partially responsible for the profound changes in the administrative structure of the city. The most significant involved a drastic reduction of Lisbon's municipal limit in mid XIXth century and a deeper transformation in the city parishes a century later. These changes caused serious methodological problems, not only because of the unawareness of the exact number of parishes at each historical moment, but essentially because of the ignorance of its limits.

The primary goal of the work I will present was to rebuild the administrative division of Lisbon for the entire XIXth century. This task was carried out using a GIS, with its CAD tools, and combining digital sources, in vectorial format, representing current parish limits and city streets [1], with historical cartography, for 1826, 1864 and 1909, where old parishes were mapped [2]. Other historical maps, old city guides, contemporary descriptions and fiscal sources were also used.

This work is part of a larger project called SIGMA, financed by the Portuguese Science and Technology Fundation, under the scientific supervision of Luís Espinha da Silveira and was carry out in the Department of History of the Faculty of Social Sciences and Humanities, New University of Lisbon [3]. Part of the methodology and some map drawing was prepared with a precious help from Sofia Lucas Martins, for who I'd like to express my thanks.

Administrative reforms in Lisbon

In the first half of the XIXth century, Lisbon had 70 parishes in a municipality that was extended for about 532 km^2. In mid century, by the 11th of September of 1852 edict, the municipality was substantially changed, reducing it practically to the urban area of the city and loosing a great number of its parishes to new municipalities: Belém and Olivais. It remained with 34 parishes in a 13 km^2 area. This new circumscription maintain itself until 1885, time when Lisbon expanded its territory up to the actual limits of the municipality. By the edict of the 18th of July, Lisbon encircled 43 parishes and an 85 km^2 area approximately. In the following year, with the edict of the 22nd of July, the capital incorporated two more parishes, which made part of the municipality until 1895, when they were separated again and incorporated in the municipality of Loures by the edict of the 26th of September. With light changes, the municipality of Lisbon that we know today is the result of this last edict.

As we can see, the XIXth century was full of changes in Lisbon's administrative circumscription but, essentially, at the municipality level. Regarding the parishes, we know that their limits were stable during the century [4]. On the urban parishes specific case, they were established in 1780 by the edict of the 19th of April, consequence of the urbanistic modifications caused by the earthquake of 1755. However, in mid XXth century this stable panorama was radically modified. The edict of the 7th of February of 1959 created new parishes and changed significantly the limits of those that remained. After this reform, the parishes gained the limits we know today, leaving behind any correspondence between the old and the new circumscriptions. It's precisely this sudden change, said in the edict as 'the most deep of all that had changed the parochial physiognomy of Lisbon' [5], which does not permit

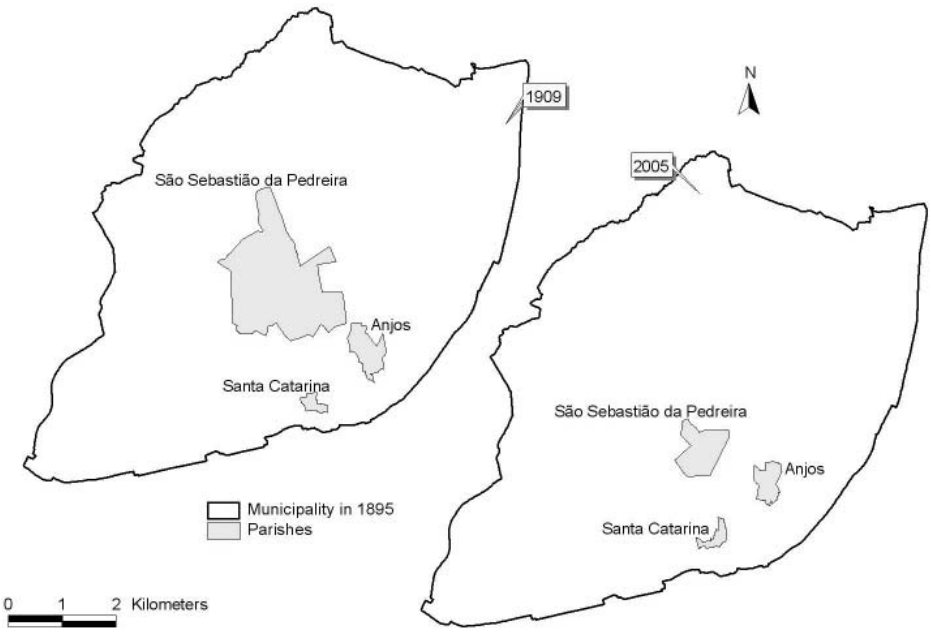


Figure 1. Parishes of Anjos, S. Catarina and S. Sebastião, in 1909 and 2005.

a retrospective reconstruction of Lisbon's parishes through the assimilation of the geographical units, methodology that was used in SIGMA for almost all the national territory [6].

Here we present only three examples of parishes changed in 1959, which illustrated very well the difficulties pointed out: the cases of S. Sebastião da Pedreira, S. Catarina and Anjos [Figure 1].

The Lisbon parishes in the XIXTh century

Knowing all these changes and their consequences, becomes obvious that for the capital could not be used the same methodology because of the risk of creating a high error level. It was necessary to start from de beginning and draw the parishes based on cartographic sources of that time. In the last years these reconstruction was tried several times [7].

1. Sources and methodology

The methodology used in the SIGMA's project has some similarities with the one used by Maria Alexandre Lousada, but we worked with technologies of CAD, GIS and relational databases. Beyond that, we did not one but three maps, that represent the evolution of the parishes in Lisbon between 1780 and 1910. Another significant difference relies on the fact that we reconstructed all the ancient parishes, urban and rural, which area was within the actual limits of the municipality.

First, it was necessary to research in different Lisbon archives looking for the required sources to do these reconstructions [8]. The objective was to find charts that had the parishes' limits represented. Of all charts found we selected five: three with the parishes' limits that we already referred [9] and other two that served as helpful material for the resolution of doubts and better identification of the city streets [10].

With the software ArcView 3.2 we used, as our work base, two digital charts, one of the roads and streets of Lisbon and another of Portugal's mainland parishes, both in vectorial format. However, they had different scales and projection systems, which compatibility was achieved by using the extensions 'Projecções de Portugal' and 'ShapeWarp', adjusting the limits between the two.

After that, all the polygons corresponding to the actual Lisbon parishes were erased, trying to obtain a work area only with the city streets, adjusted to the external limits of the municipality.

By consulting other historical charts and using old guide books [11] it was possible to identify and select, in the present road map, several of the ancient streets

Figure 2. Lisbon before the Liberalism. Municipality and parishes in 1826

that served as limit between parishes. Nevertheless, the city growth of the last two centuries, changed the physiognomy of some streets and made other disappear, creating in some cases the need for approximated drawings. In the end, we accomplished an amount of lines, selected or draw that were transformed in polygons with the 'Xtools' extension. Each polygon created represented an ancient parish, being this method adopted in all three historical charts.

However, the charts used in the first two moments, 1826 and 1864 did not represented the all municipality (which only happens in the 1909's chart); it only shows the parishes considered, at the time, as urban parishes. This fact made us face the problem of the unawareness for the 'rural' parishes' limits for the most part of the XIXth century. Nevertheless, considering the analysis of the several charts elaborated we came to the conclusion that the differences between them were minimal. This made us wandering if the same hypothesis was valid to the rest of the parishes that we already had in the 1909's drawing.

In this sense and looking forward to confirm this hypothesis, we had to try on doing a retrospective analysis of the limits in 1909. The solution was to explore fiscal sources of the city, the so called *Livros das Décimas*, which has for each parish the names of the taxpayers, their properties and it's taxes value for the first decades of the XIXth century, organised by streets. It was possible to compare these with the ones selected for the drawing of the parishes indicated in 1909 and conclude that, in most of them, the limits in this chart were coincident with the ones referred in the *Livros das Décimas* or their variation was very small [12].

In spite of the punctual errors in the definition of the limits, we concluded that, being aware of the lacking of other cartographic sources, the most correct was to assume the 'rural' parishes of 1909 to all the XIXth century. In such case, we associated them to the 'urban' parishes, acknowledge by specific sources, for each former period.

2. Lisbon before the Liberalism

To represent Lisbon in the first half of the XIXth century [Figure 2] it was used a version of the well-known Duarte José Fava's map of 1807, produced, probably, around 1820 and stored in the Military Archive [13].

In the same archive there are at least two more copies and in all of them the parishes' limits were striped in colour in the printed part of the chart. The initial work was to confront and compare the parishes in the three charts to verify if they matched, trying to check the quality of the drawings. It was possible to see that

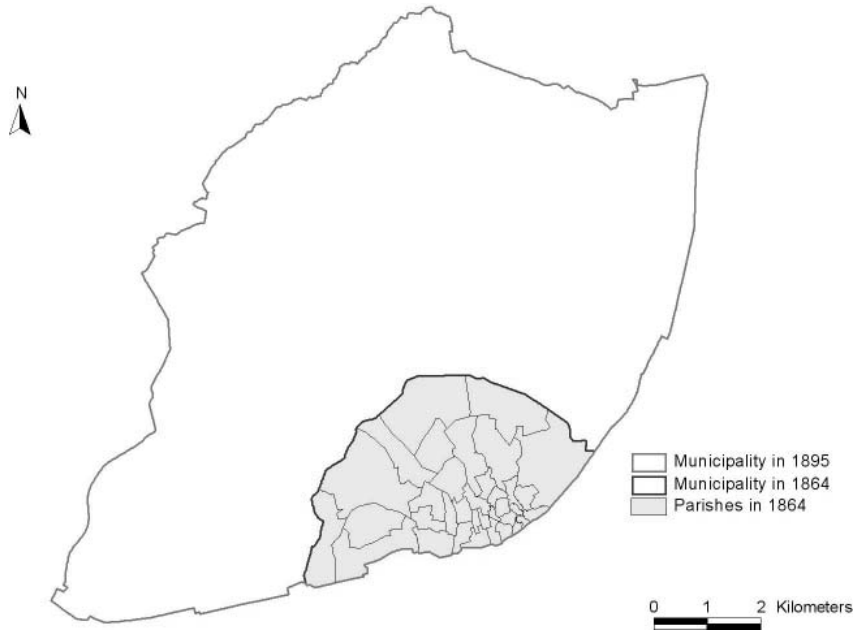


Figure 3. Lisbon in mid XIXth century. Municipality and parishes in 1864

the limits coincidence was almost perfect. Beyond that, the use of this chart revelled itself as the most correct choice by the quality that is unanimously recognised in the original survey [14].

With the *Carta Topográfica da Cidade de Lisboa, e Bairro de Belém...* it was possible to draw 39 parishes representing the end of the XVIIITH century and all the first half of the XIXTH century, since the parochial organisation of Lisbon that took place in 1780 remained until 1852, with the exception of the Belém parish creation in 28th of December of 1833 and three annexations occurred in 1835 and 1836 [15].

3. Lisbon in mid XIXth century

The city of Lisbon in mid XIXth century [Figure 3] it's the result of the work made over the *Planta da cidade de Lisboa...*, published in 1864 by Frederico Gavazzo Perry Vidal. There are some copies of this chart in the National Library, in the City Museum and in the Lisbon Geographic Society, all of them printed and with the parishes and quarter limits in several colour pencil lines.

The use of this chart considered not only its quality but also the information given once more by Vieira da Silva about his author, a civil engineer that had 'done some field work to accomplish his goals' [16].

There were made 34 parishes' drawings that corresponded to the all municipality, meanwhile reduced. The limits established for the parishes are very similar to the 1826's. Concerning the parishes that became the municipality boundary, there are obvious differences, caused by the creation of a circumvallation road, which divided some parishes intra and extra the city walls. These parishes drawing can be used between 1852 and 1885, since during these 33 years only very little changes happened caused by two parishes' annexation.

4. Lisbon in the end of the XIXth century and in the beginning of the XXth century

In 1885, Lisbon suffered a new administrative organisation, in this case, made by a new circumvallation road and the annexation of some parishes lost in mid century. The parishes' limits draw for 1885 are valid until the end of the Monarchy in 1910, except for the period between 1886 and 1895, in which the municipality still included two more parishes [Figure 4].

The draw of these parishes was based in the *Planta completa da cidade...*, published in annex of the *Anuário Comercial* in 1909 and made by Caldeira Pires. According to Vieira da Silva the 'General Caldeira Pires co-ordinated and elaborated several plans of Lisbon' [17], this one being colour printed and representing all the capital parishes limits. The chart is the National

Figure 4. Lisbon in the end of the XIXth century. Municipality and parishes in 1886

Library and it's the only example that we have knowledge of.

There were 43 parishes in 1885, 45 in 1886 and again 43 in the chart of 1909. Considering that in these years the modifications occurred at the municipality limits, with the inclusion of new parishes, and not on the parishes limits itself, we can use the same draw for the end of the XIXth and the beginning of the XXth century.

Conclusions

The methodology applied in this study could include other options, probably equally valid. One of them would be to do a digitalisation of the original historical charts and elaborate the polygons of the parishes over them. However, this would turn to be a more complex task for further compatibility between the different charts produced because they have very different scales and projections.

Nevertheless, our option, and in particular the use of the geographical information system, permitted the reconstitution of the parishes limits in Lisbon for all the XIXth century with a reduced level of error. More important, it made available a helpful tool for spatial analysis, in a vectorial format that will turn possible in the future the comparative study of multiple historic variants in the last two centuries. At the same time, it's a significant contribute for the development of the SIGMA's project since it permits to correct the lack of the system on the cartography of the main city of the country.

References

1. *Carta Administrativa Oficial de Portugal (1991)*, Original projection Hayford-Gauss Datum Lisbon, INE, 2005 and *Eixos das Via de Lisboa*, Original projection Hayford-Gauss Datum 73, Câmara Municipal de Lisboa, 1991.
2. For 1826, *Carta Topográfica da Cidade de Lisboa, e Bairro de Belém...*, scale – 1:5000, s.l., s.n., s.d., Military Archive (Engineering Corp); for 1864, Vidal, Frederico Gavazzo Perry, *Planta da cidade de Lisboa...*, scale – 1:5000, Lisboa, Lith. de Vasques & [?], 1864, National Library; for 1909, Pires, Caldeira, *Planta completa da cidade até ao limite da sua circunvalação*, no scale, Lisbon, *Anuário Comercial*, 1909, National Library.
3. The first results of this project can be consulted in http://www.fcsh.unl.pt/atlas.
4. Silva, Augusto Vieira da, *As freguesias de Lisboa*, Lisbon, Câmara Municipal, 1943, p. 20; Lousada, Maria Alexandre, *Espaços de Sociabilidade em Lisboa: Finais do Século XVIII a 1834*, Lisbon, 1995, Vol. I, p. 104; CNCDP, *Cartografia de Lisboa, Sécu-*

los XVII a XX, (exposition catalogue), Lisbon, 1997, p. 33; Silveira, Luís Nuno Espinha da, *Território e Poder*, Cascais, Patrimonia, 1997, p. 140.

5. Câmara Municipal de Lisboa, *Divisão administrativa de Lisboa*, Lisboa, C.M.L., 1960, pp. 5 to 8.
6. One detailed presentation of this methodology can be found in Silveira, Luís Nuno Espinha da, *Território e Poder*, pp. 137-140.
7. Rodrigues, Teresa, *Lisboa no Século XIX. Dinâmica populacional e crises de mortalidade*, Lisbon, 1993, pp. 40 to 51; Lousada, Maria Alexandre, *Espaços de Sociabilidade...*, Vol. I, pp. 26 to 28 and 104 to 106; Almeida, Pedro Tavares de, 'Comportamentos eleitorais em Lisboa, 1878-1910', in *Análise Social*, Lisbon, no.85, 1985, pp. 149 and followings.
8. The loot of charts representing the capital is, in most part, known thanks to the studies of Augusto Vieira da Silva and some recent expositions about historical cartography. According to those works indications we have consulted the following archives and institutions: the Military Archive (Engineering Corp), the Lisbon Municipal Archive (Arco do Cego), the National Library, the Olissiponense Cabinet Studies, the National Archive Institute – Torre do Tombo, the Portuguese Geographic Institute, the City Museum and the Lisbon Geographic Society. Silva, Augusto Vieira da, *Notícia Histórica sobre o levantamento da planta topográfica de Lisboa*, Lisbon, 1914; *Os bairros de Lisboa*, Lisbon, Imp. Lucas, 1930; *Os Limites de Lisboa*, Lisboa, Câmara Municipal, 1941; *Plantas topográficas de Lisboa*, Lisbon, Câmara Municipal, 1950 and *Dispersos*, Lisbon, Câmara Municipal, 1968; Fundação Calouste Gulbenkian, *Cartografia Portuguesa do Marquês de Pombal a Filipe Folque (1750-1900)*, Lisbon, 1982 e CNCDP, *Cartografia de Lisboa...*.
9. See reference no. 2.
10. *Planta da Cidade de Lisboa. Nova denominação dos bairros pelo Alvará de 17 de Fevereiro de 1826*, no scale, s.l., s.n., s.d., Military Archive (Engineering Corp), and Folque, Fillipe e Silva, F. M. Pereira da, *Carta Topographica da Cidade de ...*, scale – 1:5000, Lisbon, Direcção Geral dos Trabalhos Geodésicos, 1878, National Library.
11. Vellozo, Eduardo O. Pereira Queiroz, *Roteiro das ruas de Lisboa e immediações*, Lisbon, 1864, 1869, 1881 and 1888.
12. We just need to consult information regarding the non-urban parishes of the city because, for the others, we had cartography that identified their limits in different moments. It was possible to verify the limits of the parishes of Lumiar, Carnide, Benfica, Olivais, Ajuda, Alcântara and São Sebastião da Pedreira. Except for the little differences the books of 'décimas' confirmed the drawing of 1909. Archive of Tribunal de Contas, *Décima de Lisboa*, DC, 244, AR; DC, 165, AR; DC, 883, AR; DC, 40, AR; DC, 1156, AR.
13. See reference no.2. In association and as reading auxiliary to this chart it was used another one existent in the same archive from 1826. See reference no.10.
14. Silva, Augusto Vieira da, *As freguesias de Lisboa*, p. 27.
15. Silva, Augusto Vieira da, *Notícia Histórica...*, p. 28; Lousada, Maria Alexandre, *Espaços de Sociabilidade...*, Vol. I, p. 27 e CNCDP, *Cartografia de Lisboa...*, pp. 19, 33 and 34.
16. Silva, Augusto Vieira da, *Notícia Histórica...*, p. 33.
17. *Idem*, p. 44.

Broadening the scope of electronic book publishing

A comparison of commercial and public domainsoftware approaches to the development of database-driven E-resources

D. Anderson & R. G. Healey***

Introduction

Recent public announcements about the creation of very large electronic archives of printed books have highlighted the growing importance of electronic books as information repositories. However, many e-book initiatives by existing commercial publishers are very closely allied to the 'old model' of paper-based book publishing, with added facilities to locate words or phrases in the text. In fact, existing web-enabled technologies already provide extensive capabilities for a much wider range of publishing models, many of which have no obvious requirement for commercial intermediaries. Yet the adoption of these new technologies seems to be proceeding at a relatively slow pace.

This paper focuses on the creation and dissemination of e-resources, reference materials that provide significant volumes of value-added information to their internet-enabled readership. Such resources are necessarily stored in back-end databases directly connected via a suitable user interface to the web. A comparison is provided of alternative methods of developing such resources, the one based on ORACLE PL/SQL web programming, the other on use of the PHP public domain scripting tool and My SQL. Common issues arising in the course of developing such resources are also examined. The comparisons are made using two case studies. The first provides access to a digital collection of the papers of the mathematician and pioneer of computing M.H.A. Newman, the originals of which are held in the library of St. John's College, Cambridge. The second is an electronic resource on the anthracite mining industry of Eastern Pennsylvania during the 19th century. An assessment of the merits of the different approaches is given and the greatly enhanced capabilities for data linkage and flexible searching compared to standard e-book technologies are discussed.

PHP/ My SQL

In 1995 Rasmus Lerdorf developed a simple set of Perl scripts for tracking accesses to his online resume. He named this set of scripts 'Personal Home Page Tools'[1]. Over time, as more functionality was required Lerdorf developed a more feature-rich implementation in C. The resultant server-side scripting language was designed to produce dynamic web content, and it is still best suited to that task. PHP can also be used to develop full blown cross platform client-side GUI applications or to run scripts from the command line.

PHP is:

- relatively easy to learn.
- capable of generating dynamic web content very quickly.
- free (particularly advantageous to cash-strapped academics)

* History of Computing Group, School of Computing, University of Portsmouth
** Dept. of Geography and School of Computing, University of Portsmouth

– open source that is to say it is developed by the public and source code is shared.
– well supported by the development community (there are thousands of web sites offering scripts, tutorials and full blown applications.
– easily interfaced with other existing systems.
– very popular 20,478,778 Domains, 1,299,068 IP Addresses[2].
– relatively secure in operation – security announcements are shared within the community.
– efficient – demanding less written code than similar web technologies (*e.g.*. ColdFusion or ASP)

Taken together this makes PHP an excellent choice for a web programming language.

MySQL offers many of the features of a high-end database server and is capable of managing very large quatities of data. While it does not handle concurrency issues as well as Oracle, for example, it is sufficiently robust for the purposes of most web database applications and is employed in around six million web sites.[3]

ORACLE PL/SQL

The PL/SQL language is a well-established proprietary ADA-like programming language developed by the ORACLE Corporation, that contains all the usual programming constructs[4]. In addition, however, it has several important capabilities:

– It permits the seamless integration of the SQL query language and transfer of data between the database and program variables, using database cursors.
– It has automatic access to the database data dictionary, so it 'knows' about the required datatypes for a particular application
– It has standard libraries of functions and procedures that allow it to generate all the functionality of html
– It can also access other parts of the ORACLE system to display multi-media content and free text on the web
– ORACLE's webserver technology allows PL/SQL driven websites to be addressed by means of server name, port number and PL/SQL procedure name. The web site is then under the control of the PL/SQL procedure and the necessary html for individual pages is dynamically generated in response to user interactions with the site.
– PL/SQL is optimised for high performance multi-user web-based access

Equivalent functionality can be achieved by means of ORACLE JAVA programming, or indeed mixed PL/SQL and JAVA, but this is a separate topic beyond the scope of the present discussion.

Case Study 1: The Newman Digital Archive

Maxwell Herman Alexander Newman was a mathematician whose work was in the field of combinatorial topology where he greatly influenced Henry Whitehead. A series of papers by Newman on this topic between 1926 and 1932 revolutionised the field. Newman also wrote an important paper on theoretical computer science, produced a topological counter-example of major significance in collaboration with Whitehead, and wrote an outstanding paper on periodic transformations in abelian topological groups. He only wrote one book: Elements of the topology of plane sets of points (1939).

In 1942 he joined the Government Code and Cipher School at Bletchley Park. Working with Alan Turing, Newman was involved in designing and building electronic machines to break an important German cipher system, culminating in the 'Colossus' which many think of as the first electronic digital computer. Colossus was regarded as highly classified and remained completely unknown to computer scientists and the public at large until 1972.[5]

Between 1945 and 1964 Newman was Fielden professor of mathematics at Manchester University where he was the recipient of a major Royal Society grant for the purpose of developing the world's first stored-program digital computer. This was accomplished in 1948, although little or no acknowledgement of Newman's key role has so far been made. In 1939 Newman was elected a Fellow of the Royal Society, receiving the Sylvester medal in 1958. In 1962 he was awarded the de Morgan Medal by the London Mathematical Society and in 1973 he was made an Honorary Fellow of St John's College. Newman died in 1984.

The Newman Digital Archive is a digital reproduction of the Newman papers held in the library of St. John's College, Cambridge. This comprises correspondence of M.H.A. Newman with family, friends colleagues and fellow mathematicians; documents relating to Newman's father and mother; documents relating to the Government's code and cipher school at Bletchley Park and the development of the code breaking machine 'Colossus', both contemporary and post-war; Newman's will and obituaries; other miscellaneous

documents. The Newman Digital Archive is a collaborative effort between the University of Portsmouth History of Computing Group, St. John's College Cambridge and the Newman family. It aims to make the material accessible in a manner that facilitates navigation around the site and linkage between the different components of the resource.

Case Study 2: An Electronic reference work on the Anthracite mining industry of Pennsylvania

This electronic resource is a companion volume to a major interpretative study of the anthracite mining industry in the 19th century, to be published in 2006[6]. The electronic work focuses on the major mines (and their associated coal sizing and preparation plants, called coal breakers) in the Northern Anthracite Coalfield, centred around the city of Scranton, Pennsylvania. The time period covered is the 19th century. It is specifically designed for reference purposes and does not provide interpretation, as such, although the selection of content necessarily reflects in part the author's view of what was significant in the context of the wider development of the industry.

The target audience ranges from high school students engaged in projects, through to genealogists and local historians, and members of the academic community. The author has utilised his experience of working in county historical societies over many years to identify the types of questions most frequently posed about the anthracite industry. Substantial additional input about 'user requirements' has also been provided by staff from Historical Societies, local historians and colleagues from the Pennsylvania Historical and Museum Commission.

The work is organised primarily by specific mine and fifty of the most important mines in the coalfield have been selected for volume 1. Three kinds of reference data are included. The first is historical images, mostly photographs. The second is chronological information about mine developments, deployment of technological advances and major events such as mine disasters. The final type of data is largely statistical, concerned with production, employment, capacity utilisation and changing ownership. The scanned photographs have been derived from a wide variety of collections, including the author's own, and the project has benefited from the enthusiastic support of a number of organisations and individuals in this area. As a result, the archive of photographic materials available on the web site is the most comprehensive ever assembled for this period of the anthracite industry. The chronological information is even more diverse, in terms of its sources. It utilises all the research materials originally collected for the 'paper'. Sources include company correspondence and accounts, company annual reports, Mine Inspector's Reports, newspaper articles and published county histories. Much of the statistical information has been derived from the tabulations in the Mine Inspector's Reports, published after 1870, but company records and newspaper tabulations are also used extensively for the earlier period.

The electronic book is entirely driven by a PL/SQL program library, there are no static html pages. All the different types of information, including jpegs of the scanned photographs, are stored in the database and retrieved as required. The 'pages' for individual mines, can be accessed in a familiar manner, by clicking on links on the contents page or the extensive querying capability can be invoked to produce a selective contents list, which only includes pages that match the query criteria. Searches can be based on words or phrases in the chronological entries, chronological 'themes' such as mine disasters or breaker fires, value or date ranges in the statistical data, or a combination of the above. At the time of writing, consideration is being given to adding the capability to display the statistical data by year across all mines, as well as the converse, which is the current display format. In this way, the querying capabilities of the underlying database are being utilised to much greater extent than is usual in standard electronic book websites, where the focus is much more heavily towards text handling only.

For consistency, a standard format for 'pages' is used throughout. The page title is followed by a digital image or multiple images and the chronological and statistical sections continue below. The dynamic nature of the web site means that additional chronological entries, in particular, can be added by the author at any time and they are immediately reflected on the web site, as soon as the entry is saved to the database.

Implementation issues common to both case studies

a) Copyright and data security

The overwhelming majority of the 713 items in the NDA is comprised of correspondence and copyright resides with the 212 authors or their heirs. Authorship

for 61 items has not been established. Early on it was decided that material would only be made available on-line with the written permission of copyright holders. The effect of this in practice is that it will be quite some time before the whole collection may be made available. However it was possible at the outset to obtain permissions allowing 323 items (45.3%) to be included from the start.

For the anthracite e-book, the previously published materials such as the Mine Inspectors Reports are long out of copyright, and the same applies to the original photographs, so no difficulties arise in this area. However, a number of the photographs came from different archival collections, some of which derive an important component of their operating revenues from photographic reproductions. It was therefore essential that all photographs were correctly attributed to their source and sensible security measures such as digital watermarking were added, so the provenance could be traced in the event of a query arising.

b) Images and image quality

The materials held in St. John's are a mixture of originals and photocopies of variable quality, contrast levels and size. All were paper but in some cases the material is double sided and so thin that the reverse side shows through. At the insistence of St. John's no materials were allowed to leave the college and owing to the fragility of some of the documents the use of a page feeder scanner was inappropriate. It was decided to photograph the collection using a Canon EOS 300D digital camera in 8 mega-pixel RAW mode. Post-processing was carried out in Photoshop CS. For Web-based delivery, the original 8mb images are rendered down to around 50k with the inevitable loss of image quality. The aim was to produce reasonable quality images on-line while preserving archival standard images on DVD-ROM.

Provision of digital scans of photographs at a consistent level of image quality, emerged as a major challenge to the anthracite e-book project. The originals were very variable in quality, contrast levels and size. Some were mounted on buckled stiff card, while others were on lantern slides or had to be scanned through glass frames. This necessitated extensive manipulation of parameters in the scanning software used, followed, in some cases, by further image processing in Photoshop. The best of the photographs used were contact prints of outstanding quality 10' x 8' glass plate negatives from the archives of the Delaware, Lackawanna and Western Railroad. These produced excellent scans. However, for Web-based delivery, there is the obvious trade-off between quality and speed. The aim was to produce reasonable quality images printable up to about A4 without major degradation, so they could be used for accurate identification, while keeping file sizes to 70-100K. Choice of a 200 dpi scanning resolution, with moderate JPEG compression was selected as being the most suitable for operational purposes.

Comparative evaluation

Using PHP/MySQL to produce dynamic web content offers advantages of maintainability and upgradeability over static HTML. A very similar project to the NDA is the Turing Archive which comprises some 3000 static HTML pages. Using PHP/MySQL it was possible in half a day to transform this into 3 dynamic web pages with no loss of content and the addition of a searchable index. The extra functionality provided by PHP permits each page in the NDA to be dynamically linked to a photograph and web site of the copyright holder (where these are available) as well as making available associated metadata. Using static techniques this would be much harder to achieve and would give rise to serious maintenance issues.

The PL/SQL approach to web site creation and HTML generation also possesses a number of signal advantages over hard-coded HTML sites, quite apart from the obvious benefits of dynamic database linkage. The first of these is that the methods of data display and user interaction are quite generic once programmed, and therefore can be deployed for other similar reference works with minimal modification. Such works are planned for the railroad and iron industries of the same period. The second is that the complex programming logic is much easier to understand and maintain when the HTML is encapsulated in procedure calls than when it is typed out in full. Thirdly, since PL/SQL is so closely integrated with other components of the Oracle system, sophisticated processing could be undertaken behind the scenes, if required, to extend the search functionality. Examples would include complex text searching functions such as combined proximity and thesaurus searching, pattern matching across the image library, or indeed spatial proximity searching, if the geographical coordinates of mine locations were utilised. Future development plans already include

the last of these, to provide a digital atlas of the coalfield, that is linked to the existing e-book resources. Other possible extensions facilitated by the database approach include linkage from chronological descriptions to metadata indicating their source. A still further stage would involve linking through from the metadata to electronic text of excerpts from the original sources themselves, such as descriptions of the mines at specific dates, as reported by the Mine Inspectors.

Conclusions

It is apparent from the foregoing that both of the technologies discussed enable individual researchers to publish electronic resources on the web in a manner that allows flexible and powerful retrieval mechanisms to be employed. Most large funded projects with significant programming support will doubtless follow the JAVA and JSP route, although this has its own share of problems. The PHP/MySQL route has all the advantages of open source software, while the PL/SQL route, assuming the necessary institutional licences are available, gives the web author full access to industrial strength internet application development tools, coupled to the power of one of the leading commercial database systems.

Both these approaches provide an intermediate e-publishing vehicle that sits between the 'grey' and relatively inaccessible conference/working paper literature and fully-fledged peer-reviewed journal articles, whether in paper or on-line form. Unlike these publication avenues, e-publication based on dynamic links to databases dramatically reduces publication delays, increases accessibility to source or near-source materials and can be updated on an ongoing basis. It also opens up a route for the academic community to disseminate high-quality reference materials of many kinds, in a manner that we have seemingly been rather slow to exploit. The old charge of 'self-publication' can also be addressed by means of refereeing panels with oversight of content quality, and this mechanism is being employed in the second case study reported here, because of the necessary data selectivity involved. As the number of applications of the kind discussed here continue to grow, this sector of electronic publishing will begin to tax the skills of electronic librarians to keep track of metadata and cataloguing issues, but that is a topic for another day.

Notes

1 The original announcement was made on the 8th June 1995. See http://groups.google.ch/group/comp.infosystems.www.authoring.cgi/msg/cc7d43454d64d133?oe=UTF-8&output=gplain

2 Source: Netcraft http://news.netcraft.com/

3 Source: http://www.mysql.com

4 Feuerstein, S. *Advanced ORACLE PL/SQL Programming with Packages*. Cambridge, Mass: O'Reilly, 1996

5 See Randall B. 'On Alan Turing and the Origins of Digital Computers', University of Newcastle Upon Tyne Computing Laboratory Technical Reports Series No. 33 (May 1972), later published as an off-print from 'Machine Intelligence' 7, Edinburgh University Press (Nov. 1972).

6 Healey, Richard G. *From the Civil War to the 1902 Coal Strike: Recession and Resurgence in the Anthracite Coal Industry*. Scranton: University of Scranton Press, forthcoming.

Making a national census coding system internationally comparable

Trygve Andersen & Marianne Erikstad

Introduction

In Norway researchers and archivists have invested more than half a million hours digitizing historical population records. As a result the 1801, 1865, 1875 (partly) and 1900 censuses for Norway have been digitized. To simplify the statistical use of these national censuses they have also been standardized with encoding systems based on categories employed by Statistics Norway. In order to expand the application of the material even more the Norwegian Historical Data Centre (NHDC) in collaboration with the Digital Archive of the Norwegian National Archives, and the University of Bergen joined the North Atlantic Population Project (NAPP) in 2001. The project has participants from Canada, the United States, Great Britain and Iceland apart from Norway. It is a simple fact that language barriers limit the use of the Norwegian material outside of Norway. This effort to harmonize our national censuses into an international database where all samples are available in a common format with consistent variable coding and careful documentation will undoubtedly widen the user potential considerably. In this paper we wish to discuss different aspects of this process, especially the methodological ones. Our experiences may be of help to other countries that in the future will enter the dataset, like for instance Sweden. The examples will be from the 1900 census.

The recoding process

At the NHDC we store both the textual and the coded versions of the censuses in an Oracle database (see figure 1). In order to harmonize our national coded version into an international database, we had to recode our variables according to the coding system the NAPP group had agreed upon for the different variables. To make the NAPP data compatible with the existing IPUMS series of U.S. census samples, we decided to code most of the variables into the IPUMS coding system (see http://www.ipums.umn.edu/). For the occupation variable on the other hand we chose a common classification system by contextualizing the HISCO classification scheme (Leuven University Press 2002).

The simple variables like sex, marital status, and age can be converted automatically in the converting program by look-up tables for the agreed NAPP codes. The more complicated variables (see table 1) have to be dealt with in another way. They all have a high number of different strings and complex coding schemes.

Table 1. Number of different strings in the family, birthplace and occupation fields in the Norwegian 1865 and 1900 census

Field	1865 census	1900 census
Family relations	17.165	25.171
Birthplaces	41.814	77.558
Occupations	75.971	370.657
Total population	1.701.756	2.294.500

To be able to handle these variables, it is convenient to use a relational database system such as MS Access and generate a number of auxiliary tables and thereby convert the data values with the help of SQL queries. High frequency strings could be coded quite rapidly, but there are a lot of one-frequency strings that require manual interference and controls before accepting the code. When all the strings have received a proper code, the converting program runs trough the whole census and each individual receives the right codes for all variables.

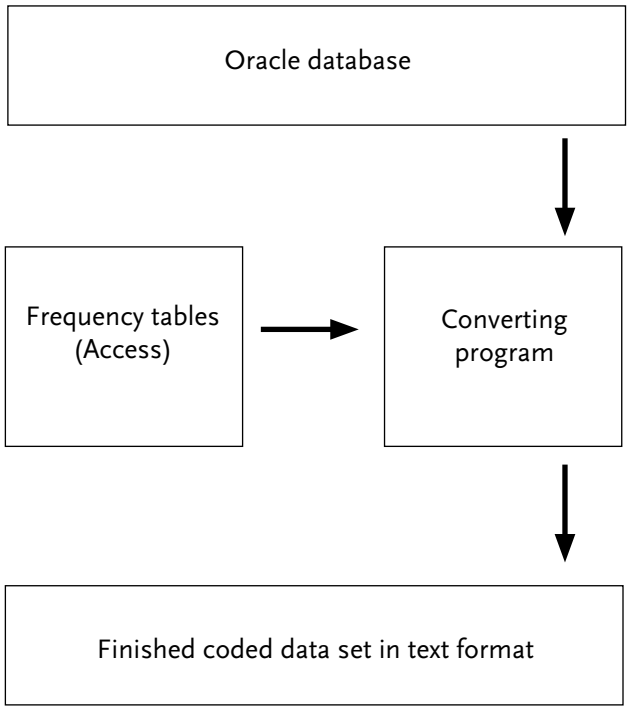


Figure 1

Place of birth

In the Norwegian censuses people's birthplace is mostly specified as a municipality within a province for the native born and as a country for the foreigners. The NHDC has used a four digit hierarchical code, where the two first digits specify the province and the two last digits the municipality within it. For foreigners only the other Scandinavian countries used to have unique four digits code, others were given a general foreign code. The US censuses on the other hand, only contain information on US state for the native born, and the country of birth for immigrants, so the birthplace variable in the NAPP dataset could only be comparable on a country level. When harmonizing the birthplace variable, it was therefore important to specify the foreigners with a proper and detailed country code. We wanted to use the IPUMS birthplace detailed codes, but reduce the original five digits code to four in order to fit it into the Norwegian system. The fifth digits mostly separate different region within a country from each other, like in Germany. It resulted in two different birthplace variables in the NAPP dataset, one for country of birth and one for Norwegian domestic locations. A problem we ran into when coding the birthplaces was that some of the Norwegian municipalities have the same name, for instance there are two places called Os, three places called Eid, and five places called Nes scattered around the country. Where there was no information about province, the only solution for some of these was to use the Norwegian country code.

Family status

The family relation variable classifies the relationship of each individual to the head of household. It sounds easy to handle, but in fact it can be very difficult, because the information in the censuses can be insufficient and the numerators did not always follow their instruction. When looking at a separate family in a census (see example 1), it can be obvious how the different members listed there relate to one another.

However, when making a compressed frequency list of this variable to be able to code efficiently, you loose this sequence information and the result of the coding can therefore turn out wrong. The same string (housewife, son, grandmother) can have different meanings in various contexts, but will receive the same code when collapsed. Another task is that when harmonizing the variable we had to split up our Norwegian family codes, since the IPUMS relate variable was more detailed than the Norwegian ones. It was also necessary to create two new codes specific to Norway. 'Retired people' (kårfolk, føderåd) receiving special benefits from farms, and 'paupers' (fattige, legsdlemmer)

Example 1. Five persons living together in a domicile in the Norwegian 1900 census

Family status	Marital status	Occupation	Birth year	Place of birth
hf (housefather)	g (married)	Fisker	1871	Sør-Varanger
hm (housewife)	g (married)	Husgjerning	1875	Finland
s (son)	ug (unmarried)	Søn	1897	Sør-Varanger
s (son)	ug (unmarried)	Søn	14.06.1899	Sør-Varanger
hfs Mor (mother)	ug (unmarried)	Vævning	1836	Finland

living with others at the expense of the municipality. The relate variable in the dataset is comparable across all countries, except Canada because they did not enumerate relationship to the head of household.

Constructed variables

As mentioned above one often wishes that it were possible to code the relate variable of each individual separately, but that will of course be far too time-consuming. Fortunately the IPUMS system has developed software that use the information in the relate variable together with data about marital status, sex, age, names and the sequence of the persons in the household to construct new variables that can give us information about family interrelationship. Among the most useful of these are pointers to own parents and number of own children for each individual if present (see example 2).

The fact that this program uses the persons' names to decide relationships, assuming surname

similarity means kinship, has made it necessary to incorporate new rules that take into consideration the Norwegian surname tradition. In the old days Norwegians were identified by their own first name and their father's first name plus the appropriate suffix for a son or a daughter (see example 3).The use of patronymic changes through the period the censuses are taken, so there would be fewer with the suffix -daughter in 1900, but the modern family names in Norway are, in fact, a product of only the last couple of generations, except among the traditional upper classes (the clergy, civil servants and the wealthy bourgeoisie). In Norway the use of fixed family names was not made compulsory by law until 1925.

Occupations

The occupation variable demanded most work. The Norwegian occupational coding system has two distinct dimensions. One dimension is the kind of trade or economic sector, that is to say the type of work a person did or the kind of business he or she was employed in. The other dimension is the place that the person's job had in the social hierarchy at the time when the census was taken. Each of the two dimensions is normally given a two-digit code. The Norwegian system is simpler than the HISCO classification scheme, so most of the coding of the occupation had to be done once more. As an example (see table 2) in the Norwegian system all the craftsmen are coded with the same code, but in HISCO all the different artisans have separate codes. I could therefore use the Norwegian system to locate them, but it was quite a

Example 2. Constructed family interrelationship variables for the family in example 1:

pernum	***momloc***	***poploc***	***nchild***	age	marst	relate	famNo	sex
1	5	0	2	29	1	101	1	1
2	0	0	2	25	1	201	2	2
3	2	1	0	3	6	301	3	1
4	2	1	0	1	6	301	3	1
5	0	0	1	64	6	501	9	2

Example 3. A typical family in the Norwegian 1900 census:

Name	**Family status**	**Marital status**	**Occupation**	**Birth year**	**Place of birth**
Peder Knudsen	hf	g	Gaardbruger S og snedker	1854	Gjerpen Brb
Anne Andersdatter	hm	g	Gaardbrugerkone	1853	Gjerpen Brb
Hilda **Peder**sdatter	d	ug	Datter	1881	Gjerpen Brb
Anna **Peder**sdatter	d	ug	Datter	1883	Gjerpen Brb
Karl **Peder**sen	s	ug	Søn	1893	Skien Brb
Andreas **Peder**sen	s	ug	Søn	1897	Solum
Knud Jørgensen	fl	e	Føderaadsmand	1814	Gjerpen Brd

job recoding them all. It would be easier to code the other way around, but we started coding with the Norwegian system long ago.

In the Norwegian censuses many of the persons have more than one occupation listed. According to the instructions to the enumerators a person's most important work should be entered first. At the time being only the first mentioned occupation has been coded, because a complete coding would take too much time. One exception had to be made in the Norwegian material, because so many in Norway live of a combination of agriculture and fishing. A new double code was therefore included into the HISCO scheme (see table 3).

Other variables

In the northern part of Norway the censuses ask for people's ethnicity. Only Canada among the other countries had the same information given, so we cooperated with them on a common coding scheme. The same applied for information about religion which we also harmonized only with Canada.

Conclusion

Working internationally is both stimulating and very interesting, because it demands a very accurate documentation of your own material that expands your own understanding of the original historical source. Using a coding scheme in a second language involves quite a lot of work especially in the beginning, because one has to study the code description carefully to be sure that it corresponds with the content of your Norwegian strings. The English speaking countries in the NAPP group have had the possibility to compare their strings directly and get help from each other in that way, but as only the Icelandic representative understands Norwegian, we are more left alone. Especially the occupation coding has required a lot of effort. Fortunately the HISCO manual gives some help in their examples in several languages after each code description. Furthermore Marianne Erikstad participated in the creating of the HISCO classification system, and therefore knows it quite well. Instructive discussions in the NAPP group have also helped solve many of our problems during these years. Looking back we feel that the work has been worth while considering the final result. A beta version of the Norwegian 1900 census was released in November 2004 (see http://www.nappdata.org/). Extracts from the census can now be downloaded and used by researchers all over the world and compared to the other countries in the project.

Table 2. Artisan occupation in the Norwegian 1900 census

Occupation	Frequency	Hierarchy	Trade	NAPPCODE	STATUS
Skomager [Shoemaker]	4194	00	21	80100	
Snedker [Carpenter]	2592	00	21	95420	
Skrædder [Tailor]	1966	00	21	79120	
Murer [Bricklayer]	1207	00	21	95110	
Smed [Blacksmith]	1013	00	21	83120	
Skibstømmermand [Shipswrights]	929	00	21	87530	
Typograf [Printers]	837	00	21	92110	
Maler [Painter]	710	00	21	93110	
Bagermester [Bager master]	543	26	21	77620	21
Væverske [Weaver]	528	00	21	75400	
Garversvend [Tanner journeyman]	426	37	21	76130	22
Malerlærling [Painter apprentice]	393	39	21	93110	23
Slagter [Butcher]	381	00	21	77310	
Blikkenslagersvend [Tinsmith journeyman]	381	37	21	87340	22
Bødkersvend [Cooper journeyman]	367	37	21	81500	22

Table 3. Double occupation in the Norwegian 1900 census.

Occupation	Freq	H1	T1	H2	T2	NAPPCODE	STATUS
Gaardbruger S [Farmer owner]	25760	01	11	00	00	61110	14
Føderaadsmand [Retired farmer]	12004	06	11	00	00	61110	11
Gaardbruger [Farmer]	5660	05	11	00	00	61110	
Gaardbruger, selveier [Farmer, owner]	5575	01	11	00	00	61110	14
Gårdbruger S [Farmer owner]	4416	01	11	00	00	61110	14
Gaardbruger Selveier [Farmer owner]	4100	01	11	00	00	61110	14
gaardbruger S og fisker [Farmer and fisherman]	2735	01	11	00	13	**61320**	14
Gaardbruger (selveier) [Farmer owner]	1900	01	11	00	00	61110	14
Kaarmand [Retired farmer]	1785	06	11	00	00	61110	11
Gaardbruger og Fisker [Farmer and fisherman]	1478	05	11	00	13	**61320**	
Gaardbrugerske S [Female farmer owner]	1253	01	11	00	00	61110	14
Gaardbruger og selveier [Farmer and owner]	1224	01	11	00	00	61110	14
Gårdbruger [Farmer]	1104	05	11	00	00	61110	
Gårdbruger, selveier [Farmer, owner]	1088	01	11	00	00	61110	14

References

Van Leeuwen, M.H.D., Maas, I., & Miles, A. (2002) HISCO – Historical International Standard Classification of Occupation, Amsterdam: Leuven University Press.

Ruggles S. et al. (1995) 'The Minnesota Historical Census Project'. Historical Methods 28(1):6-26.

Sobek M. et al. (1999) 'The IPUMS Project. An Update'. Historical Methods 32(3):102-110.

Dillon, L.Y. (2000). 'Intergrating Canadian and U.S. Hictorical Census Microdata.' Historical Methods 33(4):185-194.

Roberts, E. (2003). 'The North Atlantic Population Project: An Overview.' Historical Methods 36(2): 80-88.

Woollard, M. e. a. (2003). 'Occupational Classification in the North Atlantic Population Project.' Historical Methods 36(2): 89-96.

Computerized visual analysis of paintings

*Igor E. Berezhnoy, Eric O. Postma & Jaap van den Herik**

The paper provides insights into our efforts to develop techniques for the analysis of visual art. The AUTHENTIC project aims at creating a collection of software tools to support art experts in their assessments of the authenticity of paintings. We describe our progress on the automatic analysis of two visual features of the paintings of Vincent van Gogh: colour and texture. The colour-analysis technique is shown to confirm the generally known increase in the use of complementary colours accompanying Van Gogh's move to France. The texture-analysis technique reveals two main clusters of brushstroke shapes in a single painting. These qualifying results lead us to conclude that the use of advanced digital analysis techniques will change the way in which the authentication of visual art is performed.

1 Introduction

The assessment of paintings is largely performed by human art experts. Connoisseurship has played an important role throughout the history of art. Undoubtedly, the assessments of skilled experts are of great value to the visual arts. However, inevitably human judgements are highly subjective and prone to error. Occasionally, experts judging the authenticity of paintings made mistakes and were more then once forced to revise their opinions in the light of objective evidence. So far, objective evidence bearing on the issue of authenticity came from chemical analysis, infrared reflectography, and other examinations of the physical properties of the painting such as dendrology. In the context of the AUTHENTIC project, we attempt to develop digital techniques for an objective examination of the visual properties of paintings. The aim of the AUTHENTIC project is to develop a set of digital tools for the art expert. Using modern image-analysis and machine-learning techniques, the visual structure of (digital reproductions of) paintings can be quantified and incorporated into the overall judgement of the expert [5].

In this paper we focus on the analysis of two visual features of paintings: colour and texture. The analysis of colour is performed using a special technique that detects transitions of the two main complementary-colour pairs: red-green and yellow-blue. The analysis of texture proceeds using a tailor-made brushstroke-extraction technique to quantify the shape of brushstrokes. Both techniques are applied to digitized and colour-calibrated paintings of Vincent van Gogh. The outline of the paper is as follows. Section 2 describes the colour-analysis technique and presents the results obtained by applying the technique on 169 paintings by Van Gogh. Section 3 presents our qualifying results on the brushstroke analysis applied to the same collection of paintings. Then, in section 4 we discuss how our analysis techniques will change the way authentication and art analysis will be performed in the near future. Finally, in section 5 we draw conclusions.

* Institute for Knowledge and Agent Technology (IKAT) Maastricht University, Maastricht, The Netherlands

2 The automatic analysis of complementary colours

During his French period, Vincent van Gogh became aware of the perceptual impact of colours in paintings. He knew the effects of complementary colour pairs (*e.g.*, red-green and yellow-blue) and started to make abundant use of these colours in his paintings [1,4]. For an adequate analysis, we developed a technique that detects complementary-colour transitions in Van Gogh's paintings. In this section we describe the perceptual mechanisms underlying colour perception by focusing on opponent colours (2.1), explain the spatial filters employed to mimic these mechanisms (2.2), and present our qualifying results on the analysis of the complementary colours in Van Gogh's paintings (2.3).

2.1 Opponent colours

Colour is a mental construct (see, *e.g.*, [8,11]). Therefore, when digitally analysing colour, the brain mechanisms responsible for generating a colour experience have to be taken into account (insofar as possible). The human visual system processes chromatic signals using three types of retinal cone photoreceptors. Subsequently, the neural transformation of the signals yields an opponent-colour representation in which chromatic information is expressed in three channels: a red-green channel, a yellow-blue channel, and a black-white (luminance) channel [9,10,12,13]. The red-green and yellow-blue channels are very sensitive to complementary colour pairs. For instance, the red-green channel is sensitive to complementary colour pairs that include either red or green. Recent evidence suggests that the opponent channels arise naturally as the independent components of natural images [7]. Biological studies have revealed that individual neurons in the visual system respond to opponent colours [3].

2.2 Spatial filters

Our colour-analysis technique mimics the opponent-colour mechanism of the human visual system [13]. The technique relies on spatial filters that respond to red-green and yellow-blue transitions in the image. An example of such a filter is shown in figure 1. The figure shows a red-green filter response ('red-green value') as a function of the relative horizontal (x) and vertical (y) location of the image. The response of the filter is maximal for red-green colour transitions in the painting that match the filter location, scale, and orientation. In order to capture red-green and yellow-blue transitions

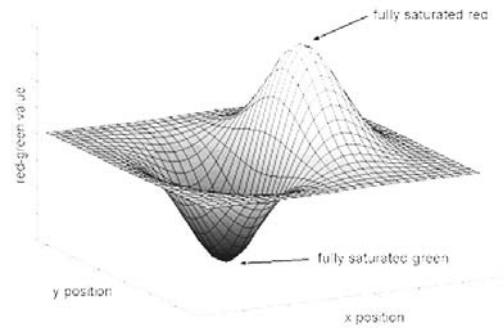


Figure 1. Example of a red-green spatial filter

at all relevant locations, scales, and orientations, a digital image of a painting is filtered by red-green and yellow-blue filters. They are centred at all image locations, and have several scales, i.e., image areas ranging in size from small (20 × 20 pixels) to large (200 × 200 pixels), and a number of orientations evenly distributed on the interval from 0 to 360 degrees. The average response of all these filters yields an (average) opponency value that indicates the amount of complementary-colour transitions in a painting.

2.3 Colour-analysis results

To quantify the usage of complementary colours in Van Gogh's paintings during his lifetime, we determined the average opponency values for 169 paintings created throughout his life. Van Gogh's paintings have been catalogued using the so-called Jan Hulsker or JH numbers [6]. These numbers correspond fairly accurately with the chronological order in which the paintings were created. In figure 2, we present the (average) opponency value as a function of the JH number for the range of about 100 (Dutch period) to about 2000 (French period). A clear transition is observed from JH1000 towards JH1400 which corresponds roughly with Van Gogh's move towards France.

Our result is consistent with the prevailing opinion of art experts and provides an objective measure of the usage of colour in individual paintings. As an example, figure 3 shows Van Gogh's painting *Landschap met bomen en vrouwelijke figuur* (JH 1848). The inset displays the image obtained by selecting high opponency values, only. The contours of the female figure are clearly visible indicating that Van Gogh employed complementary-colour transitions to emphasise contours.

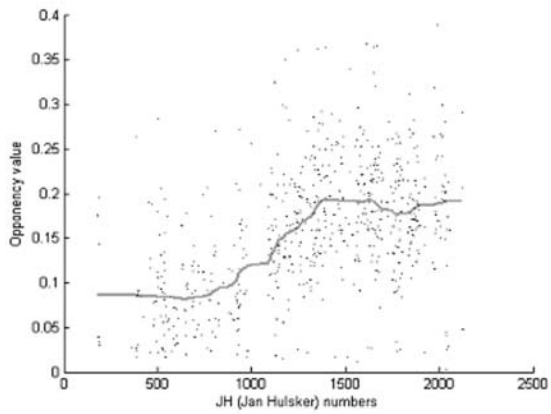


Figure 2. Average opponency value as a function of Jan Hulsker number. The dots represent the opponency values for individual paintings, the curve approximates the running average

Figure 3. The painting 'Landschap met bomen en vrouwelijke figuur' Saint-Rémy, 1889, (JH 1848). The inset displays the image obtained by selecting high opponency values only. The contours of the female figure are clearly visible

3 The automatic analysis of brushstroke texture

Texture is the second visual feature for which we developed an analysis technique. This feature reflects the statistical properties of small image regions. The statistical properties are dominated by visual contours, i.e., transitions in intensity along a contour. In Van Gogh's paintings the local texture is determined by the way he applied brushstrokes on the canvas.

3.1 Van Gogh's brushstroke alphabet

Art experts agree on the observation that the nature and distribution of brushstrokes is highly characteristic for a painter. This applies especially to Van Gogh whose brushstrokes are clearly visible due to his painting style. The brushstrokes of Van Gogh constitute a kind of alphabet, the elements of which are repeatedly used in his paintings. Examples of elements of Van Gogh's brushstroke alphabet are curved strokes, repetitive parallel strokes, and circular strokes. In an attempt to detect and quantify the occurrence of the elements of the alphabet, we started with the extraction of the shape of brushstrokes in a single painting.

3.2 Brushstroke extraction

The digital extraction of brushstrokes proceeds in two steps: (I) contour enhancement, and (II) quantification of brushstroke shape. Below, we provide a brief outline of both steps.

I. Contour enhancement. Although the brushstrokes are visually obvious for human observers, their automatic extraction is far from trivial. In order to enhance the brushstroke contours and suppress any other visual structure, we apply a circular filter to the painting. This filter enhances the parallel contours, characteristic of brushstrokes, irrespective of their orientation. Figure 4 illustrates the circular filter. The vertical axis represents the filter response; the two planar axes represent the image plane. The main parameter of the filter is the diameter of the circle which is optimised to match the average separation of the parallel contours of the brushstroke. Figure 5 shows the result of applying the (optimised) circular filter to Van Gogh's Korenveld met kraaien (JH 2117). The brushstroke contours are clearly visible.

II. Quantification of a brushstroke shape. The quantification of a brushstroke shape proceeds in three steps. Firstly, the closed contours are filled. Figure 6A illustrates a filled closed contour corresponding to a single brushstroke (or brushstroke fragment). Secondly, the closed contour is skeletonized yielding a thinned line-like representation of the brushstroke (figure 6B). Thirdly, the thinned brushstroke is fitted to an Nth order polynomial. The value of N is proportional to the complexity of the fitted curve. Figure 6C shows the result for N=3. Small irregularities are removed in this step. With these three steps, the brushstroke contours

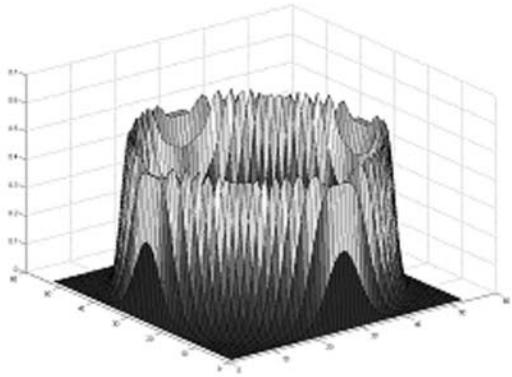


Figure 4. Illustration of the circular filter

Figure 5. The result obtained after applying the circular filter to Van Gogh's 'Korenveld met kraaien', Auvers-sur-Oise, 1890 (JH 2117).

are transformed into a compact quantitative representation, i.e., into N polynomial coefficients.

3.3 Texture-analysis results

The quantitative representation of brushstrokes enables the application of statistical analysis and learning techniques to discover painter-specific regularities. We have applied our brushstroke-extraction technique to 169 paintings of Van Gogh which resulted in over 60 thousand separable brushstrokes (and brushstroke fragments). Each of these brushstrokes is transformed into three coefficients. Figure 7 shows a plot of the distribution of the two relevant coefficients (i.e., the polynomial equals: $coef1\ x + coef2\ x^2$). The figure reveals a clear structure; the coefficients are grouped into two largely overlapping clusters. In terms of brushstroke shapes, the distribution of points represent brushstrokes ranging from slightly curved leftwards to slightly curved rightwards. Such brushstrokes correspond to letters of Van Gogh's alphabet. Currently, we are undertaking a detailed analysis of sub-clusters within the displayed distribution and within other distributions to find additional letters of the brushstroke alphabet.

Discussion

Our analysis techniques are only the start of a promising development to explore the possibilities of automatic visual examination of Van Gogh's paintings. The qualifying results on the analysis of colour and texture reveal a glimpse of the potential of digital image-analysis and machine-learning techniques. In the coming years, we will extend our techniques and results in order to obtain a full-fledged toolbox to support the art expert in his judgement.

Our techniques *support* rather than replace the expert for two main reasons. The first reason is that our techniques may reveal statistical regularities that are caused by factors other than the painting style. For instance, the distribution of brushstroke shapes may be biased by the storage or restoration of a painting. The identification of such a bias requires domain knowledge. The second reason is that our techniques are necessarily limited to low-level features such as colour, texture, and shape. High-level features such as the 'theme' of a painting are beyond the scope of present-day techniques. Their appreciation requires cultural and art-historical knowledge.

Conclusions

At present, the cultural heritage benefits insufficiently from innovations in computer science. In this contribution we provided some insights into our progress in the AUTHENTIC project. For the first time, objective visual analysis techniques are applied to the paintings of Vincent van Gogh. From the results given above and published elsewhere [2] we may conclude that the use of advanced digital analysis techniques will change the way in which the authentication of visual art is currently performed.

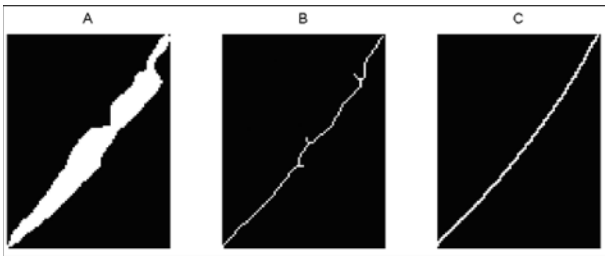


Figure 6. Brushstroke quantification in three steps: (A) filling closed curves, (B) skeletonizing the filled curve, and (C) fitting an Nth order polynomial.

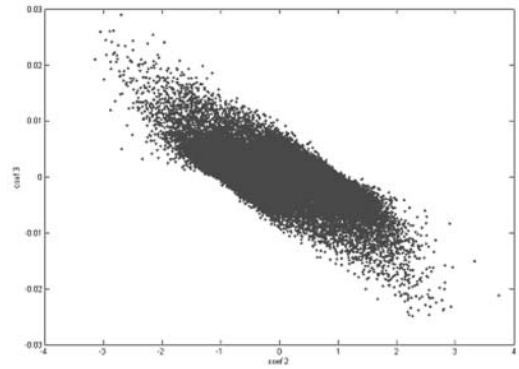


Figure 7. Illustrative results of the brushstroke-analysis technique. The plot shows the distribution of the second and third coefficient of a third-order polynomial fit to the brushstrokes in 169 paintings of Van Gogh.

Acknowledgement

This research is carried out within the Netherlands Organization for Scientific Research (NWO) Token project Authentic (grant 634.000.015).

References

1 Arnold, W.N. (1992). *Vincent van Gogh: Chemicals, Crisis, and Creativity.* Bazel: Birkhäuser.

2 Berezhnoy, I., Postma, E.O., and Herik, H.J. van den (2004). Digital analysis of Van Gogh's complementary colours. In R. Verbrugge, N. Taatgen, and L. Schomaker (Eds.), *Proceedings of the Sixteenth Belgium-Netherlands Conference on Artificial Intelligence (BNAIC-2004)*, pp. 163-170.

3 De Valois, K.K. & De Valois, R.L. (2000). Color Vision. In K.K. De Valois (Ed.), *Seeing* (pp. 129-175). San Diego, CA: Academic Press.

4 Gage, J. (1999). *Colour and Meaning. Art, Science, and Symbolism.* London: Thames and Hudson.

5 Herik, H.J. van den and Postma, E.O. (2000). Discovering the Visual Signature of Painters. Future Directions for Intelligent Systems and Information Sciences. *The Future Speech and Image Technologies, Brain Computers, WWW, and Bioinformatics* (editor N. Kasabov), pp. 129-147. Physica Verlag (Springer-Verlag), Heidelberg-Tokyo-New York.

6 Hulsker, J. (1996). *The New Complete Van Gogh.* Amsterdam: John Benjamins Publishing Company.

7 Lee, T.W., Wachtler, T., & Sejnowski, T.J. (2002). Color opponency is an efficient representation of spectral properties in natural scenes. *Vision Research*, 42 (17), 2095-103.

8 Levine, M.W. (2000). *Fundamentals of Sensation and Perception (third edition)*. Oxford: Oxford University Press.

9 Livingstone, M. (2002). *Vision and Art. The biology of seeing.* New York, NY: Harry N. Abrams, Inc.

10 Maffei, L. & Fiorentini, A. (1999). *Arte et Cervello.* Bolonga, Zanichelli.

11 Mollon, J. (1990). The tricks of colour. In C. Blakemore and M. Weston-Smith (Eds.), *Images and Understanding* (pp. 61-80). Cambridge: Cambridge University Press.

12 Wandell, B.A. (1995). *Foundations of Vision.* Sunderland, MA: Sinauer Associates, Inc. Publishers.

13 Zeki, S. (1999). *Inner Vision. An Exploration of Art and the Brain.* Oxford: Oxford University Press.

Visual object detection for the cultural heritage

*Niek Bergboer, Eric Postma & Jaap van den Herik**

The increasing availability of digital historical visual data opens up new opportunities for cultural-heritage research. For instance, modern artificial-intelligence techniques enable historians to search for visual objects, such as persons and other figurative entities, in large historical image databases. Specialised techniques may offer the ability to detect automatically particular objects from a given class of objects. This paper presents our work on the development of accurate object-detection techniques that rely on contextual cues. We provide insight into the results of our technique and discuss how cultural-heritage researchers can benefit from intelligent object-detection techniques. From these results we may conclude that in the years to come, cultural heritage research will change its research approaches significantly.

1 Introduction

The domain of cultural heritage comprises huge collections of images, photographs, illustrations, and paintings. The ongoing digitization of cultural-heritage collections makes an increasing volume of images digitally available. Nowadays, cultural-heritage researchers can search for images containing persons, scenes, and objects of their interest using textual queries. Generally, such queries are matched to the metadata or descriptions accompanying the digitally-stored images. However, often the metadata do not suffice for retrieving the information sought for. The reason is that the metadata are necessarily incomplete; it would be too strenuous to manually describe each and every object depicted in an image. In these cases, the automatic matching of textual queries to visual objects is required. For instance, when an art historian is interested in finding images depicting a certain person, or a crowning ceremony with the pope, the query 'persons' or 'popes' should return images containing one or more persons or the pope. In a similar way, more specific queries (*e.g.*, 'Willem van Oranje', 'Napoleon', or 'Pius VII') may yield images containing (or likely to contain) Willem van Oranje, Napoleon, or Napoleon's crowning by pope Pius VII.

Object-recognition techniques from artificial intelligence may support such queries for visual objects. However, the automatic recognition of visual objects is very difficult for two main reasons. Firstly, given an image it is not clear if there is a (similar) object, and where the object is located. Secondly, automatic recognition is slow.

This paper presents our work on an object-detection technique, called the Context-Based Attention (COBA) technique. COBA deals with both problems (1) by quickly selecting those regions of an image that are likely to contain the object, and (2) by validating the presence of the object using local context. In combination with off-the-shelf object-recognition techniques, COBA enables the automatic search for objects of a particular class and for their identification.

This paper is organised as follows. Section 2 describes the state of the art in the automatic searching for visual objects in image databases. Section 3 presents our context-based technique (COBA) and its implementation. Section 4 provides typical experimental results on two types of data: natural images (photographs) and artistic images (paintings). Section 5

* Institute for Knowledge and Agent Technology, Universiteit Maastricht, The Netherlands

discusses the applicability of automatic searching for visual objects in the cultural-heritage domain. Finally, Section 6 draws two conclusions.

2 Searching for visual objects

Searching for objects in images relies on two types of techniques: (1) detection techniques and (2) identification techniques. A detection technique indicates the image regions containing objects from a given class. An identification technique identifies the contents of the indicated locations. Both types of techniques are trained on instances of one or more classes in order to generalize to novel, previously unseen, instances. The main difference between detection and identification is that, for detection, the classification involves two classes ('present' versus 'not present'), whereas for identification it involves more classes or subclasses. In principle, identification of objects from a number of classes can be realised with an equal number of object detectors, each one trained to detect instances of one of the classes. Since the underlying techniques for detection and identification are highly similar, we henceforth concentrate on detection techniques.

Object detection is generally implemented as a binary pattern-classification task, i.e., image regions are classified as either belonging to the object class of interest, or belonging to 'the rest of the world', i.e., a background pattern (see, *e.g.*, [1, 3, 4, 7]). The automatic classification of visual patterns is very hard. Human observers recognize objects by combining the sensory data with prior knowledge of the appearance of objects. The computer does not have this prior knowledge and therefore has to be trained to acquire it from examples.

2.1 Training the computer to detect visual objects

Training a visual pattern classifier proceeds in two steps: (1) pre-processing and (2) classification. In the pre-processing step, the constituent pixels of the image are transformed into so-called features. Visual features summarize the information present in a large number of pixels. For instance, an image of a straight visual edge may be composed of hundreds of pixels, but may nevertheless be summarized into a few features: position of the edge, orientation of the edge, and the colours of the surfaces separated by the edge.

In object detection techniques a set of features is chosen that is well-suited to distinguish between the object class of interest and all background patterns. In the classification step, the feature values are mapped onto the two classes. The appropriate mapping is achieved by training the classifier on a large set of labelled examples. For instance, to realise a detector for faces a classifier is trained on a large set of images (instances) of both classes, i.e., instances of faces and instances of non-faces.

2.2 Dealing with false positives

Current object-detection techniques perform quite well in detecting objects of a given class (see, *e.g.*, [3, 4, 7]). However, they suffer from one main limitation. Object-like background patterns are sometimes erroneously classified as objects. Such detections are called 'false positives'. Figure 1 shows two examples from the class of *frontal human faces*. In each example, the small square pattern (left) is a face-like background pattern that was classified as 'face' by a face detector. The position in the full picture is indicated by a clearly marked square. Yet, the two patterns in question are in fact background patterns (in the upper image it is the space between two trees; and in the lower image it is a shadow pattern in the grass).

The classification errors shown in Figure 1 seem surprising. After all, humans have no difficulty in determining that the two regions in the image are not faces at all. A plausible reason for this misclassification is that humans do not only look within the square image region, but take the object's *context* into account, when determining whether a part of an image is object [2, 5].

In the next section, we present our object-detection technique that incorporates contextual information in order to improve detection reliability; the Context-Based Attention technique.

3 The Context-Based Attention technique

In this section, we describe our Context-Based Attention (COBA) technique. COBA converts the problem of object detection into two sub-problems, each of which is solved in a separate stage. We distinguish an object-detection stage, and a contextual-validation stage. Figure 2 visualizes the two stages of COBA. The first stage (the object-detection stage) is illustrated in Figure 2a. A standard object detector (comprising a pre-processing and classification step) is shifted over the image (indicated by the box with an arrow) to classify every image patch as either 'likely to contain an object' (the boxes) or 'unlikely to contain an object'.

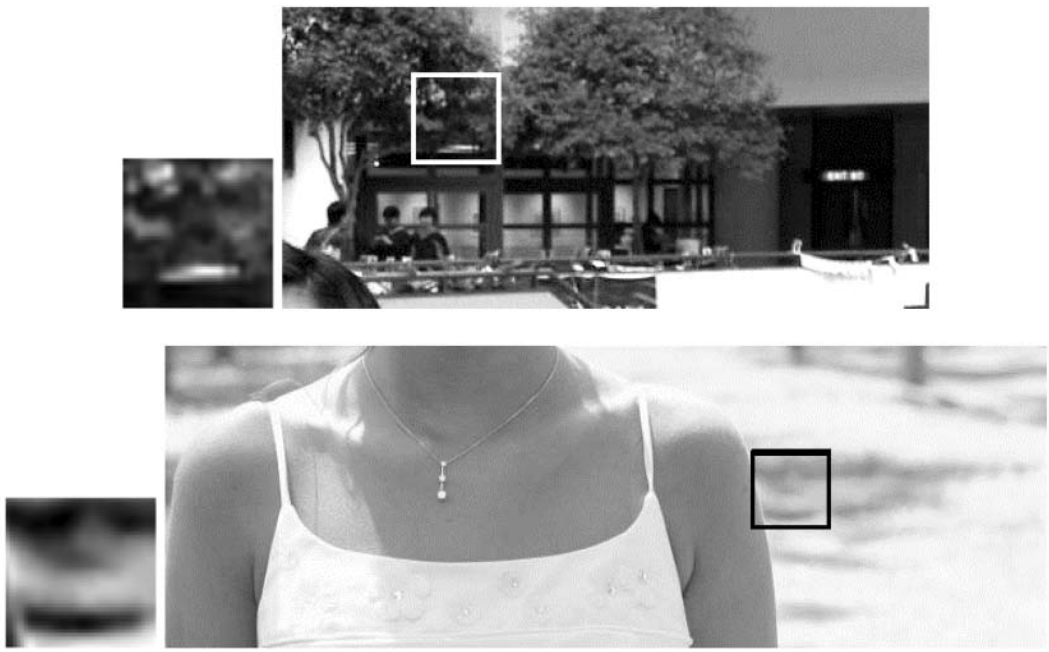


Figure 1. Examples of patterns that are similar to faces, but that are clearly not faces when viewed within their appropriate context

The object-detection stage results in a small set of image regions that are likely to contain an object.

The second stage (the contextual-validation stage), is visualized in Figure 2b. The dashed squares represent the image regions selected in the first stage. In the contextual-validation stage, multiple 'local samples' are taken in the vicinity (i.e., the local context) of the selected regions. The contents of these local samples are used to determine if a selected region does indeed contain an object.

COBA is trained on an extensive dataset of natural images containing objects of a specific class, in our case the class of frontal faces. The training set consists of an equal number of face and non-face images. A statistically-motivated procedure is employed to estimate how well COBA detects faces in novel images (i.e., images not used for training COBA).

The visual features used for training COBA are so-called Haar wavelets [6]; these encode intensity transitions in a similar manner as in early stages of animal visual systems [5]. For both the first-stage detector and the second-stage validation, machine-learning algorithms are used.

4 Experimental studies

To assess the effectiveness of COBA to the cultural-heritage domain, we applied it to two sets of images: a set of natural faces and a set of painted faces. Our set of natural faces consists of 775 images that, together, contain 1,885 labelled frontal faces ranging from 30 to 375 pixels in size. Our set of painted faces consists of 53 paintings that, together, contain 57 labelled frontal faces.

4.1 Detection trade-off

In any detection method, there is an inherent trade-off between false detections and missed detections. On the one hand, if one sets the detector threshold too low, one obtains a high detection rate, at the price of a high number of false detections. On the other hand, if one sets the detector threshold too high, one obtains a low number of false detections, at the price of a low detection rate. Detection performances are therefore often expressed in so-called receiver operating characteristic (ROC) curves. An ROC curve makes the aforementioned trade-off between false detections and missed detections visible by plotting the detec-

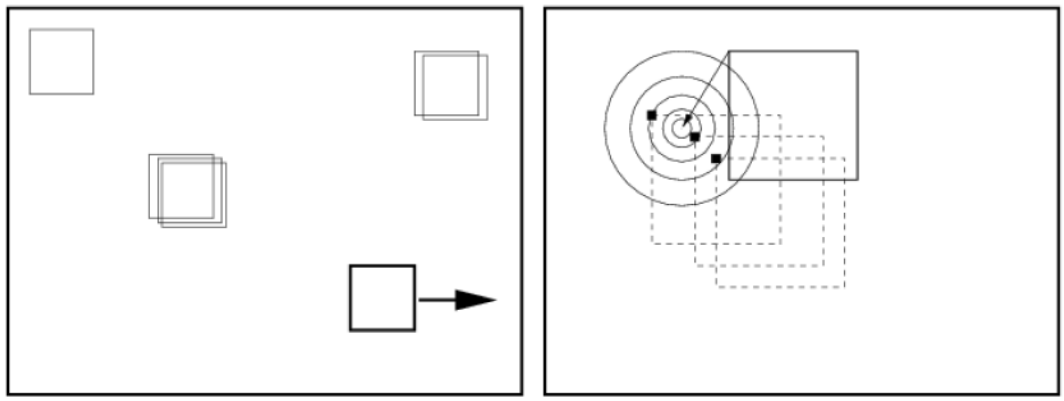


Figure 2. Schematic illustration of the two stages of the COBA technique

tion rate as a function of the false-positive rate. Good detection techniques yield curves that combine high detection rates with low false positive rates. We will illustrate the results obtained with the COBA technique in terms of ROC curves.

4.2 Object-detection results

After training on a set of labelled faces, the performance of the COBA technique on unseen images containing faces of the type under investigation, is determined. Tables 1 and 2 present the results expressed in the average number of false detections per image, for three detection rates: 80%, 85%, and 90%. Table 1 shows the results obtained on the set of natural faces, and Table 2 those obtained on the set of painted faces. The contribution of the validation stage can be computed by comparing the false-positive rates obtained with COBA (row 2, i.e., stage 1 and stage 2) to those obtained with stage 1, only (row 3). Clearly, a substantial reduction in false positives per image (FP) is obtained by the COBA technique.

Table 1: COBA's performance on the natural-faces set

	FP rate at det. rate:		
detector	80%	85%	90%
COBA	0.048	0.094	0.395
Stage 1	0.296	0.465	1.135

Table 2: COBA's performance on the painted-faces set

	FP rate at det. rate:		
detector	80%	85%	90%
COBA	0.057	0.076	0.208
Stage 1	0.223	0.893	1.204

Figure 3 displays the ROC curves for the set of natural faces (left) and the set of painted faces (right). The solid curve on the left shows that a maximal detection rate of about 90% is obtained on the set of natural faces, even when large false-positive rates are allowed. For instance, for one false detection per ten images, a detection rate of 85% is achieved. The solid curve clearly outperforms the dashed curve (representing the results for stage 1 only). The solid curve on the right shows a similar pattern. Here, the maximum detection rate is somewhat higher. Evidently, painted faces are easier to detect for COBA than natural faces. Presumably, this is due to the painter's efforts to enhance the separation between the object (face) and the background.

From these results it is obvious that the COBA technique largely outperforms the detection results with the first-stage object detector.

Discussion

The COBA technique facilitates the automatic search for visual objects and offers appropriate accuracy (low false-positive rate). In the domain of art-historical research, the usage of techniques such as COBA allows for content-based image retrieval. Figure 4 shows an example result on a self-portrait by Rembrandt van Rijn from 1669. The first (left) panel shows the original painting. The second panel shows the face candidates obtained by the first-stage of COBA; the detections are represented by white boxes. The third panel has a technical nature, it displays the regions likely to contain an object in terms of their probabilities to contain an object. Lighter shades of grey at a certain location indicate a rather high probability that an object

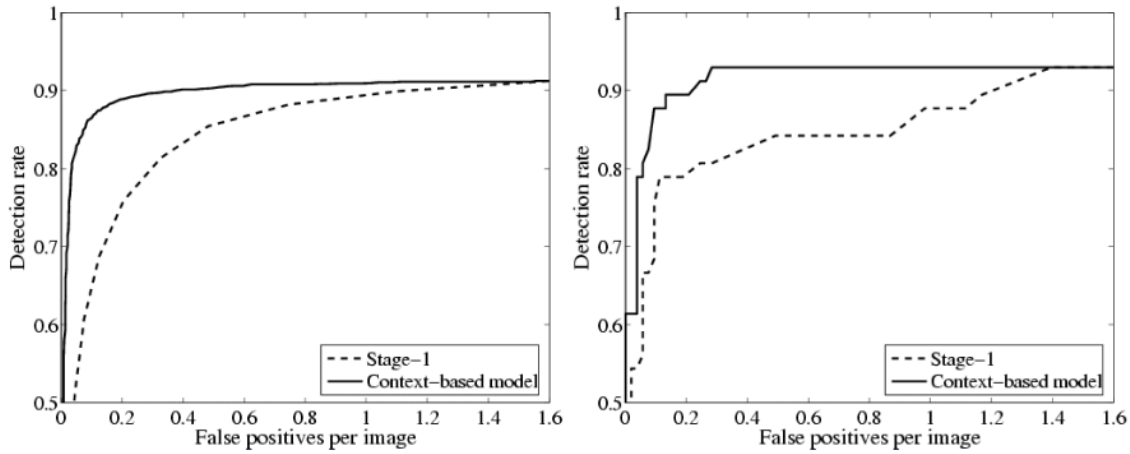


Figure 3. ROC curves for the COBA technique (solid) and stage-1 detector (dashed) obtained on the set of natural faces (left) and set of painted faces (right).

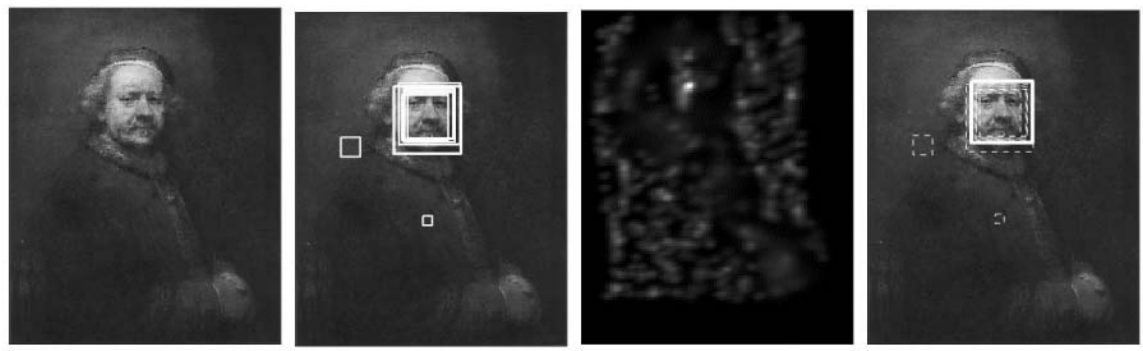


Figure 4. Detection example for a self portrait by Rembrandt (1669)

is present at that location. The right panel shows the *final detection* as a solid box; first-stage detections are shown as dashed boxes. With the present-day standard hardware, the COBA technique requires about a quarter of a second computation time per image. We admit that detecting objects at a rate of four images per second is still quite slow when searching for objects in large image databases. Future optimisation of the COBA technique and faster hardware are expected to improve the speed of detection making routine application feasible.

Conclusions

In this contribution we provided insight into the way our object-detection technique performs on sets of natural and painted objects. From our results, we may conclude that (1) it is possible to find figures, faces, and objects quite accurately in large image databases, and (2) in the years to come, research on the cultural heritage will benefit from innovations in artificial intelligence. We expect that they will change their research approaches significantly, since searching automatically for objects in large image databases may uncover unknown relations between painters and pictures.

Acknowledgement

This research is carried out within the ToKeN project EIDETIC (grant number 634.000.001) of the Netherlands Organisation for Scientific Research (NWO).

References

1 N. H. Bergboer, E. O. Postma, and H. J. van den Herik. A contextbased model of attention. In R. López de Mántaras and L. Saitta, editors, *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*, pages 927-931, Valencia, Spain, August 2004. IOS Press.

2 I. Biederman. On the semantics of a glance at a scene. In M. Kubovy and J. Pomerantz, editors, *Perceptual organization*. Erlbaum, Hillsdale, NJ, 1981.

3 E. Hjelmås and B. K. Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3):236-274, September 2001.

4 R. Lienhart, L. Liang, and A. Kuranov. An extended set of haarlike features for rapid object detection. Technical Report, Intel Research, June 2002.

5 S. E. Palmer. *Vision Science, Photons to Phenomenology*. MIT Press, Cambridge, MA, 1999.

6 C. Papageorgiou and T. Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15-33, 2000.

7 P. Viola and M. Jones. Robust realtime object detection. In *Proceedings of the Second International Workshop on Statistical and Computational Theories of Vision – Modeling, Learning, Computing, and Sampling*, Vancouver, Canada, July 2001.

Microhistory and quantitative data analysis

Heinrich Berger

Quantitative analysis of mass data is usually associated with macro-analysis and not with the analysis of very small research entities. The microanalytic approach in the social sciences was originally considered more of an alternative to the quantitative research paradigms of the 1970s and '80s. Contrary to the historical-anthropological mainstream in microhistory, though, it does not seem to be sensible to turn completely away from quantification and, in performing analysis in actual practice, to investigate only small-scale objects of scholarly research. Invoking for the purpose of illustration Siegfried Krakauer's metaphors of 'maintaining the necessary distance' to a greater or lesser degree, of 'panoramic perspectives,' and of 'the bird's-eye-view vs. the fly's-eye-view' in historical research makes it obvious that both scholarly ways of looking at things (observation from up close or from far off) have their indisputable advantages, whereby I don't want to miss the benefits of either one of these approaches.

Some historical-anthropological definitions concentrate the research perspective on 'THE HUMAN BEING'[1] or 'the concrete human being.'[2] Needless to say, human manifestations and modes of behavior occupy the focal point of interest in historical research, though it would certainly be presumptuous to declare this research to encompass the entirety of mankind. In the final analysis, it is after all only representations of human or human-designed objects, structures, relationships and patterns of behavior whose reconstructions can be investigated by historical scholars. Overemphasizing 'the general' or 'the whole' contributes more to dilution than to clarification of the historical research perspective. It might be that the great success in the field of microhistory of scholars concentrating their research perspective on individual persons[3] has even brought about a situation in which certain historians are positing that mankind itself has now become the object of research. The fact that we have thereby reached the point at which concentration on the individual does a quantum leap into a completely unmanageable undertaking becomes apparent only by stepping back and viewing things from a greater distance.

The arbitrariness of this approach is also made patently clear by the additive listing of what are said to be mankind's constituent aspects relevant to research: one such catalog cites 'actions, sufferings, perceptions, modes of behavior and basic mental states',[4] another mentions 'acting and thinking, feeling and suffering'[5] as specific objects of historical-anthropological research. The listing of these still extremely broad fields of research seems to me to be an expression of the fact that the formulation of a concrete, generally accepted research agenda has not yet been concluded. After all, there is certainly good reason to doubt whether we historians will ever be in a position to even begin to get an all-encompassing grasp of such broad areas, ones that other scholarly fields (such as psychoanalysis) often need years of intensive observations by several individuals in order to be able to come to far less widely applicable generalizations! I therefore concur with Niklas Luhmann in calling for

1 Gert Dressel, Historische Anthropologie – Eine Einführung, Vienna 1996, 25-27. The term 'HUMAN BEING' (in the German original 'DER MENSCH') appears in capitals three times on two pages.
2 Richard van Dülmen, Historische Anthropologie. Entwicklung – Probleme – Aufgaben, Köln – Weimar – Vienna 2001 (2nd Ed.), 5. A very good criticism on the concept of 'Mankind' in Historical Anthropology can be found at Jakob Tanner, Historische Anthropologie – zur Einführung, Hamburg 2004, 98-101.
3 *E.g.*: Natalie Zemon Davis, The return of Martin Guerre, Cambridge, Mass.: Harvard Univ. Press 2001; Carlo Ginzburg, The Cheese and the Worms: The Cosmos of a Sixteenth Century Miller, Baltimore 1980.
4 Gert Dressel, Historische Anthropologie – Eine Einführung, Vienna 1996, 25.
5 Richard van Dülmen, Historische Anthropologie. Entwicklung – Probleme – Aufgaben, Köln – Weimar – Vienna 2001 (2nd Ed.), 5.
6 Niklas Luhmann, Die Gesellschaft der Gesellschaft, Frankfurt 1997, 22-27.

investigation of human actions and interactions, and to dispense with talk of 'concrete human beings'.[6]

Accordingly, in the research program presented here, the object of investigation will be observed not only from close up; rather, the distance between the observer and the object will be repeatedly altered.[7] This is designed to make it possible to implement intensive interaction of macroanalytic and microanalytic perspectives and to answer the 'open question of how to build a bridge between the particular and the general, between the micro and macro levels'.[8] Generally speaking, microhistory and macrohistory address totally different issues, which is why neither of the two options can lodge a claim to universal applicability. Nevertheless, in a narrow segment of historical research, quantitatively enhanced microhistory is indeed able to bridge the gap between these two perspectives in historical research.

Although other methods and approaches have come to define the leading edge in historical research, quantification – especially on account of its solid methodological basis – is still considered one of the best-established methods in the field.[9] Much of the criticism leveled at quantification in the 1970s and '80s was thoroughly justified, and therefore could not simply be sloughed off without a response. Microanalysis offers several points of departure for arriving at a solution to the shortcomings raised in these critiques. However, that does not mean the complete exclusion of quantitative insights from historical research; rather, it means diversification of the researcher's perspective. In my research approach I undertake neither a classic microanalysis without macro-analytical links nor classic quantification without a detailed investigation of individual particulars. To the extent that it is technically and methodologically feasible, many of the issues under consideration are also subjected to a quantitative investigation. This, however, does not constitute the sole centerpiece of research; it is repeatedly, systematically enhanced by detailed analyses.

One of the strengths of quantification is still the fact that it is the best developed computer-aided analytical procedure, as well as the one that has attained the greatest relevance to research so far. Certainly its strength also lies in the thoroughly elaborated program of systematic analysis based upon solidly grounded scientific hypotheses.[10]

Formalized procedures and a stringent set of research steps – as well as, in accordance with the foregoing, the final step of inter-subjective verification – remain powerful arguments in favor of statistical procedures. This power is by no means substantially diminished by references to their inherent shortcomings such as the so-called 'loss of the individual.' In fact, I regard it as a challenge not to do without the strengths of quantification but nevertheless to respond appropriately to its weaknesses.

The positive assessment of methodological pluralism in historical scholarship has occasionally led historians to view methodological diversity as an objective worth striving to achieve even within the scope of a single research task. In going about this in actual practice, however, it is extremely difficult and often simply impossible to deploy a variety of research methods within one analysis – whereby this applies especially to the attempt to overcome the antagonism of microanalysis and macroanalysis. The central problem here is the theoretical link-up of the various issues under investigation and the practical integration of the analytical results of these different research strategies. One of the few fields of research that offers relatively favorable opportunities to overcome methodological limitations is the systematic analysis of small groups of persons. The small number of individuals under investigation makes it possible to analyze them both quantitatively (above all, descriptively-statistically) as well as qualitatively. The linkage takes place via the individual persons, who are both part of overall structures while also existing within their own personal networks. This means that the sequence of analytical

7 Siegfried Kracauer, Geschichte – Vor den letzten Dingen, Frankfurt 1971, 103, 109, 123, 124f.
8 Jürgen Schlumbohm, Mikrogeschichte – Makrogeschichte: Zur Eröffnung einer Debatte, in: Jürgen Schlumbohm, Mikrogeschichte – Makrogeschichte. Komplementär oder inkommensurabel?, Göttingen 2000, 28.
9 Here Lutz Raphael speaks about 'more precise methodical coverage' (in the German original: methodisch präziserer Absicherung'): Lutz Raphael, Geschichtswissenschaft im Zeitalter der Extreme. Theorien, Methoden, Tendenzen von 1900 bis zur Gegenwart, München 2003, 175.
10 Vgl. Jürgen Kocka, Quantifizierung in der Geschichtswissenschaft, in: Heinrich Best / Reinhard Mann (Eds.), Quantitative Methoden in der historisch-sozialwissenschaftlichen Forschung, Stuttgart 1977, 4.

steps must proceed both deductively and inductively, whereby, once again, the diametrical opposition of macroanalysis and microanalysis is overcome.[11]

During the process of gathering data, the analyst must make sure that the random samples are truly representative. Needless to say, though, when data are quantified within the scope of a microanalysis, the limited number of cases makes it unreasonable to demand that the source material satisfy this claim to representativeness. Small investigative entities can also lead to highly atypical historical conclusions;[12] therefore, the effort must be made to utilize comparative methods to integrate sources under investigation into the respective historical context.

If a combination of different types of sources is necessary, then control variables should also be introduced in order to check the compatibility of the data. Another way to combine two types of sources is the combination of two data sets by means of a (nominative) record linkage with the help of a databank system.[13] Especially when performing databank-aided microanalyses, the linkage of several sources is often very useful.[14]

Research design

In contrast to methods used in mainstream microhistory, this work will not attempt to completely dispense with quantification; rather, it will make an effort to repeatedly vary the distance between the scholarly observer and the object of research (variation of scale, level switching).[15] The core of this investigation is a very intensive, multi-perspective analysis of a relatively small group of persons – that is, a few houses and its changing, growing population. Moreover, the results of these analyses will not be presented in isolation (classic local study); instead, they will be placed into a frame of reference of broader fields of research. In this study, though, this will be accomplished not only by raising 'big' questions and hypotheses that are then tested in the narrower entities to be investigated, but also checked against a few basic social-structural parameters in the small entity as well as on a larger scale (individually-related mass data).

In selecting a small neighborhood as typical example for a larger group of people or for a larger field of research, it is a priori impossible to lodge a claim of representativeness. Therefore, the comparison with larger-scale, quantitatively utilizable entities is highly beneficial, especially for a series of tests of representativeness.

For this project, I have chosen a Viennese housing complex whose individual buildings were not torn down during the period of time under investigation and which displayed a quantitatively relevant proportion of Jewish residents (almost a third) at a relatively early stage (1857). In connection with this aim was the fact that 'Jewish infrastructure' (of a religious, cultural and commercial nature) was available to a considerable extent in the complex itself and in the contiguous quarter. The housing complex is bordered by avenues and side streets, and the adjacent square underwent a major status enhancement from the site of a prison to a marketplace ('Karmelitermarkt') during the period of time under investigation.

All of these circumstances seem to me to be very typical, important and – for the conduct of the investigation – desirable characteristics of the research entity I have chosen. At the same time, though, it must be pointed out that these assumptions are initially to be considered unproven hypotheses that must be subjected to testing over the course of the analysis. Therefore a systematic comparison with other research entities is necessary again.

---

11 A good description of the differences between micro and macro history and of the ways overcome this antagonism can be found in: Jakob Tanner, Historische Anthropologie – zur Einführung, Hamburg 2004, 110-112; or in: Siegfried Kracauer, Geschichte – Vor den letzten Dingen, Frankfurt 1971, 123.

12 Richard van Dülmen, Historische Anthropologie. Entwicklung – Probleme – Aufgaben, Köln – Weimar – Wien 2001 (2nd Ed.), 104-105.

13 See *e.g.* Wilhelm H. Schröder, Historische Sozialforschung: Forschungsstrategie – Infrastruktur- Auswahlbibliographie, in: HSR, Supplement I, 1988, 18; or Heinrich Berger, Record Linkage with Multi-Lingual Sources in Early Modern Bohemia, in: Peter Denley / Mathew Woollard (Ed.), The Sorcerer's Apprentice: Case Studies, St. Katharinen, 129-136.

14 See *e.g.* Hans Medick, Weben und Überleben in Laichingen 1650-1900, Göttingen 1996; David W. Sabean, Property, production and Family in Neckerhausen 1700-1870, Cambridge 1990; Jürgen Schlumbohm, Lebensläufe, Familien, Höfe. Die Bauern und Heuerleute des Osnabrückischen Kirchspiels Belm in proto-industrieller Zeit, 1650-1860, Göttingen 1994.

The key questions of this investigation have to do with, on one hand, the social, economic and religious structure of the households and the people who make them up, as well as, on the other hand, the personal relationships among the residents of the housing complex. One of the essential methodological problems of this research program is (as outlined above) the linkage of the micro-perspective and the macro-perspective.

The greatest challenge in going about this is the practical matter of linking up the various different issues and, following the analysis, setting the results of the two research strategies in relation to each other. It is precisely this sort of systematic analysis of small groups of people (in this case, the residents of a housing complex) that is among the few fields of research for which this boundary-transcending approach is appropriate. Due to the small number of cases (a maximum of 697 persons), the sample can be subjected to both quantitative (descriptive-statistic) as well as qualitative analysis.

The starting point of this investigation is a census list from the mid-19th century.[16] In order to obtain an overview of the object of this research, the census information is subjected to a very detailed descriptive-statistical analysis. Then, a 'thick description'[17] of the single buildings (construction; owner; commercial businesses and institutions located in the houses, etc.) and households (structure; occupants, etc.) will be performed.

As soon as the results of this investigation are available, I have to face the challenge of not only interpreting them using classical microanalysis but also systematically analyzing them within broader historical contexts (Vienna as a whole; comparison with other cities in Austria and Central Europe). However, since this investigation started with a census list, no theoretical bridges or leaps are necessary, since the same descriptive-statistical investigation of larger comparable groups from the same type of source (data from the Vienna Database on European Family History)[18] can be used for the synopsis.

As far as the practical analysis is concerned, this means not only that, for example, the information the people gave about their origins can be quantified (counted, aggregated, weighted and correlated) and compared with larger data samples, but also that the migration paths of the individual persons in the entity under investigation can be precisely described, and then compared with and related to their respective social milieu. Which persons (Jews or non-Jews of which age) came from which familial and occupational backgrounds at which times and from which geographical areas to Vienna? In which buildings (Jewish or non-Jewish owner; location of the house) and household configurations (core family or extended family household; Jewish or non-Jewish head-of-household; cohabiting coworkers and/or subtenants) did they live? Which occupations (traditional or new) did they pursue? Which personal relationships did they enter into? Furthermore, in individual cases, we also have access to information as to whether they later moved away again and, if so, where to, as well as in what social framework (move to another address, emigration, flight, Shoah) the subsequent migration took place.

In light of the fact that the starting points of the analysis of these migration biographies are census data of a small entity here, I also have the opportunity to establish whether the migration pattern identified in the present case can be classified and allocated to a larger group and/or one of which the selected sample is representative.

Thus, proceeding from a detailed, intensive analysis of a small, urban object of investigation and by means of a wide-ranging comparison, it is possible to make generalizations about the history of Jews in Central Europe and, moreover, to do so over a timeframe extending from the Revolution of 1848 to the time of National Socialism.

15 Jakob Tanner, Historische Anthropologie – zur Einführung, Hamburg 2004, 111.
16 Sources from the Vienese City Archives, 'Konskriptionen 3. Reihe Leopoldstadt 1857'.
17 Clifford Geertz, Thick Description: Toward an Interpretative Theory of Culture, in: Clifford Geertz, *The Interpretation of Cultures. Selectes Essays,* New York, 1973, 3-30.
18 Samples Leopoldstadt, Schottenfeld, Sechshaus and Hernals of the the 'Vienna Database on European Family History'.

Advancing digital scholarship using EDITOR

*Peter Boot**

Introduction

The EDITOR program is an annotation tool for scholarly digital editions. 'Annotation' should be understood in a wide sense: it includes text commentary, categorization of text fragments according to any typology, creating references to other works, and the creation of connections between text fragments.

Modern scholarly editions of literary, cultural or historical texts usually distinguish between a source format (probably XML) and (multiple) presentation formats (probably web pages). Scholarly annotation should refer to the source files, not the web pages, as the latter are just transitory representations of the scholarly objects that need annotation.

Annotations that unambiguously refer to locations in the XML source file of an edition, are suitable for exchange with other researchers basing themselves on the same edition. Display of the annotations can also be integrated into the web edition itself, if the provider of the web edition feels they represent a worthwhile enhancement to that edition.

EDITOR is being developed at the Huygens Institute. Work up to now has concentrated on developing the annotation component. Web display of annotations and integration into existing digital editions will be taken on at a later stage.[1]

Examples in this paper will come from an investigation into a 17th century emblem book, Otto van Veen's *Amoris divini emblemata* (Antwerp 1615).

The problem

Well-made digital editions offer superior search and navigation facilities, but their true potential is wider. Digital editions can support the day-to-day working processes of the humanities scholar, in writing, in publishing and in teaching. EDITOR was conceived to assist in the annotation of scholarly digital editions: in the definition, creation, maintenance, display, manipulation, exchange, and archiving of annotations. EDITOR is based on the premiss that, for many humanities scholars, research consists to a large extent in note-taking, in making lists, in categorizing and commenting on text fragments. Many of these annotations (again, in a wide sense) are just working material and are discarded at a later stage; others will be used in publications: in articles, reports, books, and syllabuses. Increasingly, one will expect this material to be available for digital exchange, display and manipulation.

The Huygens Institute is the Royal Netherlands Academy of Arts and Sciences institute for scholarly editing and intellectual history. The Institute creates scholarly editions of works from the early middle ages up to the present. We want our editions to be superior tools for research for our colleagues at universities and elsewhere. Now that we are moving to create digital editions, we will want to assure their usability in research. EDITOR is meant to contribute to that usability, and thus to the creation and feasibility of text-

*Huygens Institute, The Hague, The Netherlands pboot@xs4all.nl

1 EDITOR is under active development at the time of writing (May 2005). For the latest developments, see http://www.huygensinstituut.knaw.nl/editor. EDITOR is being developed in close cooperation with the Royal Netherlands Academy of Arts and Sciences I&A development group, whose help we gratefully acknowledge. The author also wishes to thank his colleagues at the Huygens Instituut, especially Hanneke van Kempen and Herman Brinkman, for the time and effort spent on EDITOR development.

2 Peter Robinson, 'Current issues in making digital editions of medieval texts – or, do electronic scholarly editions have a future?', Digital Medievalist 1.1 (Spring 2005).

based digital scholarship. Despite recent suggestions to the contrary [2], EDITOR development assumes new scholarly editions will overwhelmingly be digital. The true potential of the digital edition will only be realised once the edition can serve as a vehicle for digital scholarship.

We anticipate that institutions that host digital editions, such as the Huygens Institute, will facilitate the use of EDITOR on these editions. Scholars will be able to annotate the editions using EDITOR, and if the scholar agrees, the hosting institution may decide to offer access to his or her annotations from within the digital edition.

Related work

There are many programs that offer annotation facilities. Especially researchers in linguistics and the social sciences have developed annotation software. The social science programs for 'content analysis' often come with facilities for automated text analysis, statistical analysis and sometimes text mining software. EDITOR differs from these programs mainly in that it was conceived to annotate digital editions, i.e. publicly available XML files.

What sets EDITOR apart from much of the text analysis software developed for use in the humanities is, again, EDITOR's focus on sharing annotations in a web environment, but also our assumption that most of the annotation will be done by hand. We expect most of the interesting observations about edited texts will be made by humans, assisted by computer programs, not the other way around.

A program that often comes up when discussing EDITOR is Annotea, developed at the W3C[3]. Annotea is server software for storing and retrieving annotations. It needs an annotation client for actually creating and displaying annotations. We felt existing Annotea clients were mainly targeted at simple web pages and did not provide an adequate platform for annotating large edition files. We might have used Annotea itself for storage of the annotations, but felt there was little point in using it outside of its intended scope. Storage of annotation data in a server component would also raise the privacy and data ownership concerns we discuss below.

A very interesting annotation product is APE, the 'Assistant for Philological Explorations', developed by Dieter Köhler[4]. APE specializes in annotation creation, not display. Köhler has spent much thought on the issue of canonical reference systems as used in referring, for instance, to locations in the Bible or in the works or Aristotle. Using APE, the researcher's annotations can use these canonical references rather than pointers to a single digital edition. This is clearly an important issue, which EDITOR does not expect to address.

As EDITOR is still very much in development, positioning it among similar programs would be claiming either too much or too little. We believe EDITOR, once completed, will be unique in simultaneously being all of the following:

– an open source program
– suitable for the creation, manipulation and display
– of diverse annotation types
– on XML based digital editions
– which allows for displays integrated into the edition
– where annotations will be visualized at multiple levels and in multiple modes

We will be the first however to acknowledge that much remains to be done.

Using EDITOR

Editor will facilitate annotation of scholarly editions. The present author is using it right now in research into an emblem book digitized at the Emblem Project Utrecht, Otto van Veen's *Amoris divini emblemata* (Antwerp 1615)[5]. In this religious emblem book, the figures of the human soul and divine love are represented in symbolically significant situations. Relevant questions are (to me): whose opinions are being represented in this book? Who is convincing whom? Who gets to speak, who is supposed to listen?

EDITOR allows me to annotate the emblems and text fragments: who speaks, who says what, who is told to listen. Do the texts use the first, second or third person? Even in its present state, without the web display component, EDITOR is already helping me to organise and navigate through the material.

Of course, this is only one of many subjects that can be studied using EDITOR. A literary historian can use

3 See http://www.w3.org/2001/Annotea/

4 See http://www.philo.de/ape/

5 Emblem Project Utrecht (EPU): see http://www.let.uu.nl/. The edition of Amoris divini emblemata is available at http://emblems.let.uu.nl/emblems/html/v1615front.html.

EDITOR to study the use of figures of speech. A theologian might study the respective roles of virtue and grace, or an art historian the depiction of nature in the emblem prints.

At the Huygens Institute, we also anticipate using EDITOR in a study of the internal coherence of a large medieval miscellany, in annotating the realia in a Renaissance ode to the city of Amsterdam, and in a study of the changing arrangement of verse in the books of 19th century Dutch poets.

The program

Interface

As stated, work up to now has concentrated on creating the annotations. Work on web display will begin later. The main window of the program offers a view of the CSS-styled edition XML. A secondary window on the left displays a tree view of the XML document, which can be used to navigate through the document. A raw XML view is also available.

Annotations can be created on text selected in the presentation window or on nodes selected in the tree view. The program will prompt for the annotation type and then present a window where the annotation text and other data can be entered. The presence of an annotation is marked in the main view using brackets, in the tree view using bullet nodes. A separate subwindow shows all existing annotations (or a subset of them, based on type or value). Annotation types are user-defined. An associated colour is used for ease of visual distinction between types. Figure 1 shows the main elements of the user interface.

Data model

The user can create multiple sets of annotations on a single edition. Each annotation set is stored in a separate file. The data format used for the annotations is RDF. Figure 2 shows the data model being used. An annotation set can contain annotations of multiple types. Each annotation can contain multiple fields, as defined at the annotation type level. For fields that are

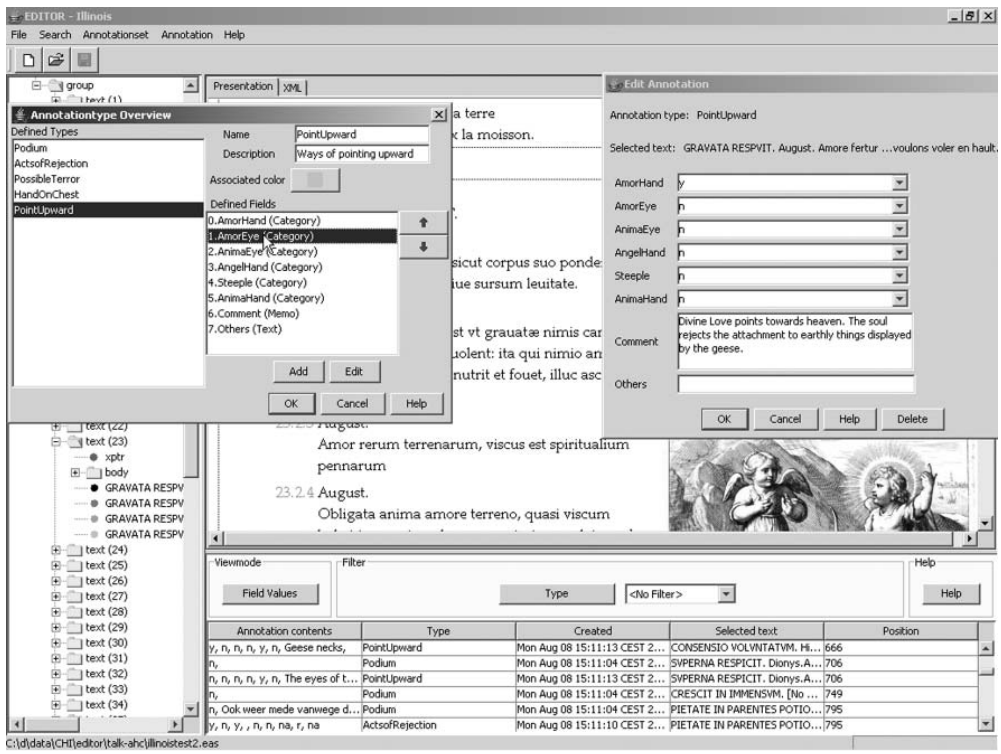


Figure 1. EDITOR user interface. The annotation displayed shows how a single annotation can consist of multiple fields. The Annotation type overview window displays the corresponding Annotation type definition.

used to categorize text fragments, the system remembers the values used before in order to facilitate data entry.

The annotations refer to fragments in the edition by describing their begin and end points. The points are stored using an xpath-expression to identify a document node and an offset within that node.

Input fields within the annotations get their name and type from the Annotation Type Field defined at the Annotation Type level. Allowed types are (at present): category, text field, URL field, memo field and related annotation field. A related annotation field refers to an existing annotation and hence allows annotation of relations between text fragments.

EDITOR requirements

Development of EDITOR has been guided by an analysis of the requirements that an annotation tool should fulfil. General requirements are:

1. The software should be available as open source.
2. The output and storage formats should be based on public standards.
3. It should be possible to share annotations and annotation display definitions.

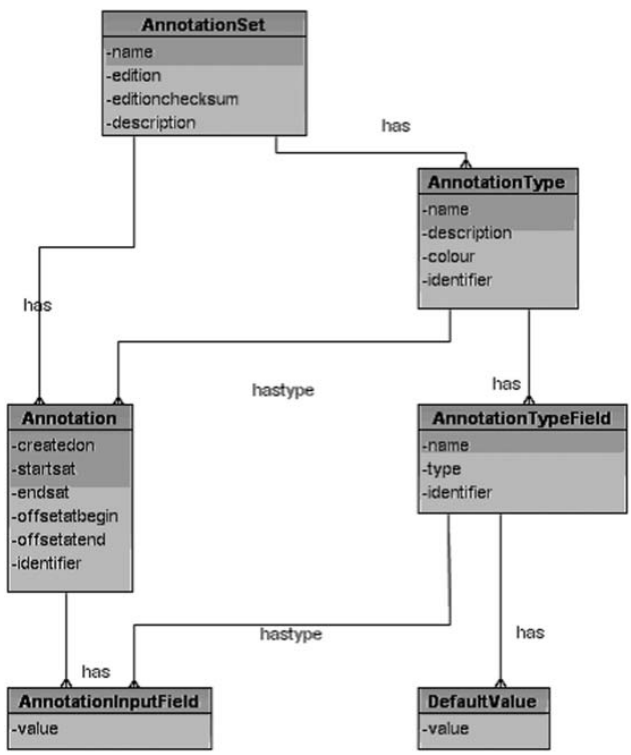


Figure 2. EDITOR data model

4. The software should work with 'any' modern digital edition.
5. Results should be accessible from the digital edition, and offer entry into the digital edition

More specifically, basic requirements for the creation of annotations are:

a. The annotations should be safe. The creation of annotations represents a considerable investment of time; there should be no doubt about the continued availability of the annotations, whatever the fate of the EDITOR program.
b. The researcher should be in control. Annotations may be experimental or personal. The researcher should feel no concerns about the privacy of work that he/she has not explicitly decided to share with others.
c. The annotations should unambiguously identify locations (text fragments) in a publicly available edition. The annotations should not, therefore refer to web pages (which will change as technology and interface design develop). Instead, they should refer to a presumably durable edition source format.
d. The edition itself should remain untouched. The edition source file is the result of probably years of careful editorial work. Storing the annotations inside this file introduces the possibility of corrupting the editorial work. It would also create a bottleneck in the annotation process, as multiple scholars may be working on the same texts.
e. The researcher should not need special ICT expertise in order to work with the annotation toolset.
f. The annotation types should be user-defined. EDITOR should be useful to a wide range of humanities (sub-)disciplines, and no amount of predefined annotation types will be sufficient to satisfy scholars from all fields. (And of course, it is in the nature of scholarly work to want to introduce new distinctions, rather than to rehash the old ones).

If these criteria are met, the annotations will be in a suitable format for display in (some sort of) conjunction with the edition to which the annotations refer. The connection may be one-way only (hyperlinking from the annotation display to the edition), but in other cases will go both ways, making available the annotations from the existing edition.

Requirements for the annotation display are:

a. There should be a filter facility for the annotations (by annotation type, by annotation value, by annotated fragment location).

b. There should be a sorting facility for the annotations (again, by annotation type, by annotation value, by annotated fragment location, perhaps by annotated fragment type).
c. Annotation display should be optionally based on a tree view of the edition's XML structure, making it immediately clear which text fragments carry which annotations.
d. Annotation display should be optionally aggregated at higher edition hierarchy levels (for instance counting all annotations of a certain type that occur on verse lines and showing their number on the stanza level).
e. It should be possible to introduce a break level or break locations; thereby facilitating side-by-side display of edition sections and their annotations. Side-by-side display of sections (for instance: different versions of poems; earlier and later sections in a manuscript) makes it possible for the researcher to *e.g.* compare properties (vocabulary, style, etc.) between sections.
f. In annotation display, it should be possible to switch the document hierarchy to an alternate hierarchy for instance: from a hierarchy based on books, chapters and section to a hierarchy based on the physical document structure, in quires, pages and columns).
g. It should be possible to base annotation display on a joint display of multiple hierarchies.
h. It should be possible to visualize annotation counts (in any display unit) using diagrams. If for instance each dialect word has been annotated, it should be possible to display in a diagram the number of dialect words per stanza.
i. It should be possible to store annotation display definitions. Researchers should be able to store multiple display definitions for use by visitors. Visitors should themselves have the option of modifying these definitions and storing them for their own private use on later visits.
j. The main platform for annotation display should be the Web. The Web is presumably the place where the digital editions themselves will be available, and increasingly, the platform of choice for the distribution of all scholarship.
k. Annotation display should offer a clear API to the world, in order to facilitate the display of annotations in external contexts.
l. Annotation display should be, wherever possible, hyperlinked to, and from, the web edition.

Some of the implications of these requirements will be discussed in the next section.

Some issues

Writing down requirements is the easy part of software development. Making the decisions about how to implement them and actually implementing them is the hard part. This section will discuss some of the decisions we made about the implementation.

Local vs. central annotation entry

It is one of EDITOR's central requirements that annotations can be shared with the world. It was clear from the beginning that annotation display should be web-based – though not necessarily always on a publicly accessible web site. Though a strong case can be made for web-based annotation entry too, we decided otherwise.

The advantages of web-based annotation entry would be: sharing of software functionality between annotation entry and display components, accessibility from anywhere in the world, no requirements for local installation.

However, we felt the advantages of local processing outweighed these. Perhaps the most important of these advantages relate to EDITOR's trustworthiness for the researcher. The researcher's privacy should be guaranteed: his possibly experimental annotations should not be exposed to preliminary scrutiny by others. Perhaps more importantly, his intellectual property should remain his own. Storing annotations in a central server would expose them to risks that PC-based storage would avoid. A second aspect of trustworthiness is the annotations' guaranteed availability. Web server based storage is beyond the control of the individual scholar. The scholar can however save and backup files on his own computer.

A second important argument for local processing is the need for EDITOR to handle large edition files and large numbers of annotations. We anticipate edition files of several megabytes in size, carrying tens of thousands of annotations. It is true modern web programming techniques reduce the need for wholesale transfer of large files, and high speed internet access is nearly universal. Still, we felt the dedicated processing capability of a personal workstation was our best bet for the sub-second response times that we wanted to achieve.

Supported edition formats

The Huygens Institute will create its digital editions using TEI/XML.[6] Beyond its HTML presentation, there will always be a permanent and durable source format that is the mother format from which the HTML representation derives. This durable source format will use XML for its syntax and the TEI tagset for the definition of the edition contents.

The diverse nature of the Huygens Institute's edition projects however, precludes prescribing the use of a single DTD or XML schema. An edition of a medieval manuscript tradition needs other elements than one that edits a modern book of verse. EDITOR cannot therefore assume the use of a single DTD.

Even if the Huygens Institute itself were to decide upon a single DTD, we still would not want to limit the use of EDITOR to editions that use this 'Huygens DTD'. EDITOR was meant to be potentially useful for every scholarly edition. What EDITOR does need, however, is an XML file. The file does not need to be in TEI format. However, if EDITOR encounters n-attributes (used in TEI for text, page and line numbers) it may use these to facilitate navigation.

Local edition display

The decision to annotate the XML source format of a digital edition implies EDITOR needs to display the edition XML to the researcher. Again because of the differences between the edition XML files, it should be possible to use different display definitions for different editions. What we need therefore is a flexible mechanism to apply style to XML files.

It was clear to us we didn't want to develop our own standard for XML display, which left us with a choice between CSS and XSLT[7]. XSLT, while very powerful, would not just style, but also transform the XML, and thus break the connection between the displayed, annotated text locations and the corresponding locations in the original XML file. Which is why we chose to use CSS for defining the display of the edition within EDITOR.

For those editions that are based on TEI, the CSS style sheets to be used will likely be very similar. No two editions are the same, however, and especially things like text numbers, line numbers and columns numbers – very essential to ease of navigation – depend very much on the texts being edited. Using EDITOR on a new edition will often therefore entail creating a custom CSS style sheet. We anticipate these style sheets will usually be created by the institution hosting that digital edition, not by the individual researcher.

Technically, the edition display uses JRex, a Java wrapper around the (Mozilla) Gecko rendering engine. The style sheets therefore can use Mozilla-specific CSS-extensions[8].

Conclusion

On of the more important considerations in the design of EDITOR has been its usability beyond the editions of the Huygens Institute. A number of scholars and scholarly institutions from the Netherlands and elsewhere have expressed their interest in the results of EDITOR development. Some of those have expressed some form of readiness to participate in the program's development.

We do not expect problems in EDITOR's usability to other institutions in the creation of annotations. All that is needed for using EDITOR is an XML file and a CSS style sheet. However, problems may arise in the integration between annotation display and the (existing) web display of the digital edition. How will the existing web edition of, say, *Amoris divini emblemata* 'know' that there are annotations to be displayed with certain text fragments? Moreover, if it 'knows', how is the process that generates the web edition from the XML source to take account of that information?

This is one of the larger issues which EDITOR development will have to tackle in the near future. We still welcome persons and institutions willing to help in thinking about and building this and other EDITOR components.

6. The Text Encoding Initiative (see http://www.tei-c.org/) proposes guidelines for the creation of digital editions of literary, historic and cultural material.
7. CSS: Cascading Stylesheets (http://www.w3.org/Style/CSS/). XSLT: eXtensible Stylesheet Language Transformations (http://www.w3.org/Style/XSL/).
8. Among which is the use of XBL (eXtensible Binding Language) to introduce some extra formatting (images, tables) and behaviour (hyperlinks).

CDWA lite for Cataloguing Cultural Objects (CCO): A new XML schema for the cultural heritage community

*Karim B. Boughida**

Metadata standards: a brief typology

Before I begin to describe our project, I think it is important to specify and differentiate the standards involved in cataloguing and indexing cultural heritage objects. The descriptive metadata experts classify data standards in four families:

– Data content
– Data value
– Data structure
– Data format.

Data content standards are the rules that guide you in filling a particular metadata field. [RLG, p.4]. Examples are: *International Standard Bibliographic Description* (ISBD), *Anglo-American Cataloguing Rules 2* (AACR2), and *Cataloguing Cultural Objects* (designated here as 'CCO').

Data value standards are the thesauri, controlled vocabularies, and authorities that provide terms and other values with which to populate a metadata field. [ibid. RLG, p.4]. Examples include the *Art & Architecture Thesaurus* (AAT), *Union List of Artist Names* (ULAN), Library of Congress Subject Headings (LCSH), *Thesaurus for Graphic Materials* (TGM), ICONCLASS [note 1], etc.

Data structure standards specify the categories and organization of metadata elements. A data structure standard constitutes the framework that holds the relationships and hierarchy between elements. Examples include *Categories for the Description of Works of Art* (CDWA), the Visual Resources Association (VRA) Core Categories, MARC [note 2], Dublin Core, etc.

Data format standards (or what I call data 'technical expressions,' since format can be an ambiguous term) are the technical expression or encoding of data in a file or a database table. XML schemas and document type definitions (DTDs) are examples of data formats. The CDWA Lite XML schema that I will describe in this paper is one such technical expression. Other examples are the Dublin Core (DC) schema, the MARC format, MARCXML, the VRA XML schema, the EAD (Encoded Archival Description) XML DTD, the MODS (Metadata Object Description Schema) XML schema, and so on.

The genesis of CCO

CCO is a data content standard designed to set the rules for cataloguing cultural materials and their visual (including digital) surrogates. For the purposes of CCO, a cultural object is defined broadly as any object that has an artistic or cultural and historical value. This covers museum objects, photographs, archeological artifacts, etc. It is not intended to cover archival collections at the group level, nor bibliographic items, which have their own data content standards – Describing Archives: A Content Standard (DA:CS) or General International Standard Archival Description (ISAD(G)) for archival collections, and AACR2 for bibliographic items. But this does not mean they could not intersect. For example, DA:CS could be used at the collection level for describing objects with a common provenance, and at the item level, CCO could be used. For books and book-like objects, MARC can be applied, but for example an interesting frontispiece or the prints contained within the book could be cata-

* Karim Boughida, Senior Information Systems Architect, Getty Research Institute / The J. Paul Getty Trust (http://www.getty.edu), 1200 Getty Center Drive, Suite 1100, Los Angeles CA 90049-1688 USA, kboughida@getty.edu

logued using CCO principles. This leads to the obvious questions: How to embed the relationships and hierarchies in different systems, and how to present the final product to end-users so it makes sense and can be interpreted? These are open questions. All of the stakeholders (cultural heritage institutions, standards organizations, consortia, vendors, end-users) are trying to build bridges and protocols to facilitate such a concept; among those initiatives is the NISO (US-based National Information Standards Organization) metasearching group [note 3].

CCO was conceived in 1999 by a group of VRA members, and took shape as a formal project in 2001. It was intended to address the lack of data content rules to accompany data structure standards such as CDWA and VRA Core, and to enable cultural heritage institutions to share and contribute descriptive metadata. CCO benefited from the work of earlier initiatives and projects such as the Museum Educational Site Licensing project (MESL), the Visual Resources Information Online Network (VISION) and Record Elements for Art and Cultural Heritage (REACH), and was heavily informed by – indeed, we might say designed to complement – data structure standards like CDWA and VRA core. With regard to data value standards, CCO strongly encourages the usage of established vocabulary standards and thesauri such as LCSH, the AAT, ULAN, the TGM, and so on.

From CCO to CDWA lite

CCO as a data content standard and CDWA as a data structure standard needed a technical structure in which they could be expressed. Data structure standards such MARC, Dublin Core, and MODS could have been used, but they all have their own limitations, mainly because they were addressing the needs of specific communities (i.e., the library and Web resources communities), and not specifically the needs of the cultural heritage and art information communities. In a 'community meeting' that took place at the Getty Center in Los Angeles in November 2004, participants began talking about an XML schema that would be the 'container' for CCO. Some were in favor of a neutral schema (that is, a cross-community schema for museums, libraries and archives). Some were in favor of creating a profile within an already established schema, such as MODS.

In 2001, before CCO had taken any kind of definitive shape, the Getty Research Institute had created an XML DTD (a DTD being the oldest schema format) for cataloguing unique objects from their Special Collections. A working group was formed, composed of members representing different departments of the institute (Standards and Digital Resource Management, Special Collections Cataloguing, Collection Development [the curatorial department of the institute], and Library Information Systems. It was my task to the design the XML DTD that we called 'Getty VRA' [Baca] This DTD was based on the VRA Core Categories (which are directly based on CDWA), with the addition of some CDWA subcategories, such as Provenance. The Getty VRA XML DTD became the *de facto* internal data format standard for the Research Institute's descriptions of unique objects. Some of the pieces of this 'local' DTD were later incorporated into the forthcoming official VRA Core XML schema.

So why hadn't a CDWA schema been created earlier, especially since the Getty is the official custodian of that particular data structure standard? Before 2004, CDWA was considered to be a data structure standard, not a format for the technical expression and delivery of data [note 4]. CDWA could have been implemented using several XML DTDs/schemas, or even relational tables. We were in favor of allowing multiple interpretations of how fields are named, arranged, etc. but when CCO began to become a concrete project, the need for an agreed-upon technical vehicle became clear. Since CCO is based upon CDWA (or rather, 'takes for granted' the CDWA elements), it made sense to have a CDWA schema, but considering that CDWA has more than 300 categories and subcategories, we decided that a 'lighter' version would be more practical and feasible to implement: thus was born 'CDWA Lite.' That said, the real impetus from the Getty's point of view was the desire to create a standards-based way of contributing Getty records to union catalogs like RLG Cultural Materials and ARTstor. Sharing records in the form of union catalogs is an old tradition in the library and museum worlds. MARC has long been the main technical format for library materials, but there was no comparable standard format for cultural objects or visual resources.

Kenneth Hamma, Executive Director of Digital Policy and Initiatives at the Getty, was in favor of also developing a new way of contributing records to union resources. Instead of manufacturing records or doing custom exports over and over again per the requirements of content aggregators like the former Art Museum Image Consortium (AMICO), RLG, and ARTstor, Hamma believed that it would be much more efficient

if the formatting and export of data to be contributed to union resources were done once, in a standards-based way, and the final product could be routinely exposed as harvestable records via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Murtha Baca, a member of the CCO editorial team and Head of Standards and Digital Resource Management at the Getty Research Institute, was the advocate for the creation of a new standard XML schema based on the CDWA core categories.

The members of the working group to whom Hamma gave the task of developing a standards-based way of contributing Getty records via OAI-PMH harvesting initially discussed several technical solutions, including the usage of Metadata Encoding and Transmission Standard (METS) or Moving Picture Experts Group 21 – Digital Item Declaration (MPEG-21 DIDL) [Van de Sompel, Digital Library] as a metadata 'wrapper' for the descriptive elements plus the accompanying images. Those two emerging standards are addressing the very real problems of packaging and sharing complex digital objects, but in light of the available resources and the quite short timeline, we decided not to pursue those technologies for now. The group will continue to follow the work of Herbert Van de Sompel et al.[Van de Sompel, Resource Harvesting].

To ensure that the would be viable and would have a 'real-life' implementation, a partnership was initiated between the Getty and ARTstor. It was agreed that records and images for paintings on display in the Getty Museum, and records and images of European tapestries from the Getty Research Institute's Photo Study Collection would form the initial testbed for the project.

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

OAI-PMH was an obvious choice for making our records harvestable. OAI-PMH had already proven its efficacy as a standard in the digital library world for exposing and sharing metadata. The main players in OAI-PMH are data providers (DPs) [note 5] and service providers (SPs) [note 6]. For this project, the Getty would be the DP, and ARTstor would be the SP.

For the purposes of this project, 'records' were defined in their generic meaning – specifically in this case, mostly descriptive metadata about objects in the Getty collections. OAI-PMH deals exclusively with descriptive metadata, not with the resources (digital images, multimedia files, and so on) related to the object being described. A strong debate commenced when we started thinking which descriptive metadata element set to use, since the usage of Dublin Core had been considered mandatory in OAI-PMH up to that time. But the OAI community and especially the Digital Library Federation (DLF) OAI Best Practices had already begun moving towards the usage of richer metadata element sets than Dublin Core. So, the use of a set of metadata elements specifically designed to describe cultural objects (for which Dublin Core is neither designed nor particularly effective) as the descriptive metadata was deemed to be the best course of action. It may take years before the cultural heritage community at large will be able to access a critical mass of objects described in this schema, but we believe that there was a compelling need for a new standard for contributing cultural heritage records and for resource discovery. The SPs (including vendors) we approached with this idea showed enthusiasm for supporting this schema, provided that it would be accepted and enriched by the cultural heritage community at large.

Eventually, the goal is to also enable our metadata records and related images and other media to be harvested by commercial search engines like Google and Yahoo. As of this writing, Yahoo now is harvesting metadata records from a major SP, OAIster [note 7], and Google is extracting the URLs in the identifiers tags from the same SP.

For the technical implementation of OAI, we decided to use OCLC tools

(Java-based and open source) [note 8]. The DLF OAI Best Practices group is working on facilitating the choice of technical tools by building a decision tree based on functional requirements. At the DLF OAI Best Practices meeting in San Diego in April 2005, I suggested the creation of a benchmark/matrix of what the various vendors' tools offer.

Other pending OAI issues are defining and describing a 'set.' This is a somewhat vague term used to denote groups of records that are exposed for harvesting in various phases. A 'set' of harvestable records is not necessarily the same as an entire discrete 'collection' of records for a cohesive group of materials (*e.g.*, the Getty Museum paintings or the Research Institute tapestry collection). The DLF group is working on this issue. Building an XML schema to express sets has been proposed in order to enable interoperability between record contributors (DPs).

We also need to define OAI identifiers, which are not to be confused with persistent identifiers or PIDs. OAI data repositories should conform to the XML schema for the OAI Identifier Format. Use of this schema will mean that the repository identifiers are globally unique within the OAI namespace [Warner]. The implementation and management of a persistent identifier infrastructure is parallel to the defined OAI project. OAI does not depend on a persistent identifier scheme to function properly. Meanwhile, having true persistent identifiers in the future will supercede the need for an OAI identifier. A persistent identifier is global and unique. The usage of 'handle technology' as an infrastructure, coupled with a good naming scheme (not necessarily based on handles), seems to be a good strategy. As of this writing, the timeline of the OAI-PMH/CDWA Lite project is supposed to be completed (i.e., the records exposed in the CDWA Lite schema, along with their related images, and successfully harvested by ARTstor) by September 2005.

Harvesting resources (digital images)

Until the experimental research on harvesting resources such as digital images and other media via OAI-PMH is mature, we will use a traditional approach. The resources or images will be uploaded to a File Transfer Protocol (FTP) server and the descriptive metadata will reside on the OAI server. We are still pondering about the best way to synchronize the harvesting of records and images; as of this writing, we are working with ARTstor, our partner on this pilot project, on the best way to manage this workflow. A workflow engine could be customized. An XML Rich Site Summary (RSS) feed could be set up so that whenever an image is uploaded it will send a message to the SP.

A separate group at the Getty is working on the specification for the images to be harvested. It was decided that we would not put TIFF images on the FTP server, since the impact on the bandwidth would be enormous. For now, we are going to use a high-resolution JPEG format:

Pixel X dimension Average 4224 / Pixel Y dimension 2897
X resolution 1000 / Y resolution 1000
Colorspace sRGB (Standard Default Color Space for the Internet)
Average Size 1.5 to 2 MB
Bit depth 24
Color true RGB

Then what will happen to the images harvested? For images in the public domain, the Getty is in favor making everything available for the «public good». Kenneth Hamma has stated that:

> Instead of asserting intellectual property rights in images of public domain works as nearly every art museum does now, it is argued here that publicly and pro-actively placing these images in the public domain and clearly removing all questions about their availability for use and reuse would likely cause no harm to the financial position or trustworthy reputation of any collecting institution and would demonstrably contribute to the public good. [Hamma]

Methodology of building the CDWA lite schema

Our past experience with building the Getty VRA DTD and being exposed to digital library schemas like METS, Dublin Core, and others was very helpful as it provided us with a proven methodology for building a schema. A schema construct has two main components: a specification (sometimes called a data dictionary) and the XML Schema Definition (XSD), expressed in XML. Why use XML in the first place? For reasons of interoperability, and to be following a technical standard. XML allows users and contributors in the future to be able to read and interpret data without the need for specific tools. Also, XML is human-readable [note 9]. We also believe that XML and other markup languages (vs. relational tables) in general are well suited to describe hierarchical relationships among data elements, and they are natively Web-ready.

So, the first step in developing the CDWA Lite schema was to develop the specification or data dictionary. Several staff from various Getty departments, together with technical and metadata experts from ARTstor, developed this document, which contains the following: version history, purpose of the schema, a numbered list of the elements, element tags, descriptions, repeatability, requiredness, data values, tagging examples and display examples (see the sample in Appendix 1).

The idea was to have a good compromise between CDWA (a very rich set of metadata elements) and Dublin Core (a relatively poor set of elements). The document (and the resultant schema – see Appendix 2) has two main sections: descriptive metadata (19 top-level elements) and administrative metadata (3 top-level elements). The last section is needed in order to make it possible to adequately process the records.

Descriptive metadata (CDWA-based) cannot carry elements such as links to the resource, rights, etc. We made a clear distinction between the record ID, the record info ID (ID of the metadata, not the record), and the resource ID (in general the surrogate or image ID, not the metadata).

We followed several best-practice principles to ensure that the metadata elements and sub-elements were 'wrapped' correctly:

– Usage of XML namespace cdwalite to avoid collision with other schemas that have the same tag name.
– Most of the top elements are wrappers. The goal is to let XML processors know which tag belongs to which set of elements, *e.g.*:

```
<cdwalite:titleWrap>
<cdwalite:titleSet>
<cdwalite:title pref='preferred'>Po
rtrait of Maria Frederike van Reede-
Athlone at Seven Years of Age</cdwal-
ite:title>
<cdwalite:sourceTitle>J. Paul Getty
Museum. Handbook of the Collections.
Los Angeles: J. Paul Getty Museum,
1991. <cdwalite:/sourceTitle>
</cdwalite:titleSet>
</cdwalite:titleWrap>
```

Here sourceTitle and title tags go together inside titleSet.

– Have extra wrappers for future use in a more robust CDWA schema
– Usage of Set to separate instances and allow sub-elements to belong to one parent. In the example above, we could have one or more additional titleSet elements.
– Authorities, *e.g.* the AAT, are embedded in attributes:

```
<cdwalite:objectWorkType
termsource='AAT'>cartes-de-visite</cd-
walite:objectWorkType>
```

– We have a good mechanism for handling related works (parent-child-sibling-ascendant-descendant relationships).
– We are still working on repeatable vs. non-repeatable and required vs. not required issues. SPs fear that if there are too many required elements, records will be rejected. We feel that if a record lacks what we think is required, it is not considered a good CDWA Lite record. The missing data should be supplied, or other, less rigorous schemas may be applied in these cases.
– The schema follows the basic CCO principle of separating display and indexing elements.
– Basic principle in CDWA, CCO and VRA: separation of work and image or surrogate.

Mapping collections

As mentioned above, two collections were chosen for this project: Getty Museum paintings on public view, and images of tapestries from the Photo Study Collection of the Research Library of the Getty Research Institute. The first collection lives in a collection management system (relational database). The second one lives in a flat-file system with linking capabilities.

The goal is to do a delimited ASCII export, since these systems do not support XML, and then use custom tools to create well-formed, valid XML documents. We will use eXtensible Stylesheet Language Transformations (XSLT) to repurpose or clean the data as necessary.

Conclusion

CCO and its technical corollary, CDWA Lite, are a potentially powerful way of describing and disseminating descriptive metadata about cultural objects. Our goal in developing the CDWA Lite XML schema was to break new ground by creating a technical standard (combined with appropriate data structure and data content standards) that will enable description, contribution, and discovery of metadata and related resources about cultural objects. By helping to shape a new standard, our hope is to serve the cultural heritage community at large and to prepare that community to take advantage of the new realm of the digital world where the barriers between data, metadata, and media resources are becoming increasingly blurred. This new standard is compliant with and uses existing digital library standards and protocols; it could also be used as a building block for more complex METS or MPEG-21 DIDL digital objects in a federated Web context.

Ours is still very much a work in progress, and we certainly have not exhausted all of the possibilities of the standards and technical protocols that are now available to us. We recognize that – like any pilot project – the CDWA Lite/OAI-PMH harvesting project has its limitations, and may seem to treat communities as if they were monolithic organizations. But it is

our hope that with the help of these very communities – both national and international – this initiative will continue to evolve and mature, so that we can reach consensus on a standards-based way to express and share information on our collections that will one day become a routine part of the workflow of cultural heritage organizations.

Acknowledgements

I would like to thank the other members of the team that developed the CDWA Lite XML schema: from the Getty (Murtha Baca, Erin Coburn, and Patricia Harpring); and from ARTstor (Ameer Ahmed, Emerson Morgan, Dustin Wees, William Ying). CCO is a VRA-sponsored project, with funding from the Getty Foundation, the Andrew W. Mellon Foundation, and the Digital Library Federation.

References

Baca, Murtha (2002). *Documenting Visual Culture at the Getty: Contributing to the Toolset. VRA Bulletin* 29 (4), 49-60

Hamma, Kenneth (2005). *Public Domain Art in an Age of Easier Mechanical Reproducibility*. CNI Project Briefing: Spring 2005 Task Force Meeting. Retrieved May 01, 2005 from http::/www.cni.org/tfms/2005a.spring/ s/PB-hamma-public.html

RLG Cultural Materials Initiative, Description Advisory Group (2005). *Descriptive Metadata Guidelines for RLG Cultural Materials.* Retrieved May 01, 2005 from http::/www.rlg.org/en/pdfs/RLG_desc_metadata.pdf

Van de Sompel, Herbert (2003). *Digital Library Architecture based on MPEG-21 DIDL, the OAI-PMH and the OpenURL*. CNI Project Briefing: Fall 2003 Task Force Meeting. Retrieved May 01, 2005 from http::/www.cni.org/tfms/2003b.fall/ s/PB-digital-sompel.html

Van de Sompel, Herbert; Nelson, Michael L.; Lagoze, Carl; Warner, Simeon (2004). *Resource Harvesting within the OAI-PMH Framework*. D-Lib Magazine. December. 10, 12, doi:10.1045/december2004-vandesompel. Retrieved May 01, 2005 from http::/www.dlib.org/dlib/december04/vandesompel/12vandesompel.html

Warner, Simeon. (2005, February 15). *Problems with 'OAI Identifiers' section*. Message posted to listserv [Oai-best-practices] Retrieved May 01, 2005 from http::/comm.nsdl.org/pipermail/oai-best-practices/2005-April/000340.html

More on CCO

Main web site: http::/www.vraweb.org/CCOweb/ (Retrieved May 01, 2005)

Lanzi, Elisa (2004). *Cataloguing Cultural Objects: new guidelines for descriptive cataloging*. Art Libraries Journal, 29(4), 26-32

Baca, Murtha (2005). *Descriptive Cataloging: an Essential Piece of the Digital Puzzle*. Conference proceedings and solicited papers. Chicago: ALCTS Publications. In process.

Notes

1 Iconclass is a subject specific international classification system for iconographic research and the documentation of images (www.iconclass.nl)

2 MARC is used here as generic for MARC21, UNIMARC, FINMARC, etc.

3 URL NISO MI http://www.lib.ncsu.edu/niso-mi/index.php/Main_Page

4 MARC, for example, is both – it defines the elements or categories for a bibliographic record, and also provides the technical 'container' for expressing, exchanging, and delivering the data contained in such a record.

5 OAI-PMH Data Provider a data provider maintains one or more repositories (web servers) that support the OAI-PMH as a means of exposing metadata.

6 OAI-PMH Service Provider a service provider issues OAI-PMH requests to data providers and uses the metadata as a basis for building value-added services.

7 URL OAISTER http://oaister.umdl.umich.edu/o/oaister/

8 URL OCLC OAI www.oclc.org/research/projects/oai/default.htm

9 XML is not particularly user-friendly. The goal is to make it readable not necessary understandable. <cdwalitedisplayEdition>46/500</cdwalitedisplayEdition> is human-readable whereas '1000111101001011' is not.

Appendix 1

CDWA Lite: XML Schema Content for Contributing Records via the OAI Harvesting Protocol

---

3. Element: Display Creator

Element tag: <cdwalite:displayCreator>

Description: The name, brief biographical information, and roles (if necessary) of the named creator or creators in the design and production of the work, presented in a syntax suitable for display to the end-user and including any necessary indications of uncertainty, ambiguity, and nuance. If there is no known creator, make a reference to the presumed culture or nationality of the unknown creator.

Non-repeatable

Required

Data values: Formulated according to data content rules for creator display in CCO and CDWA; may be concatenated from the Indexing Creator elements, if necessary. The name should be in natural order, if possible, although inverted order is acceptable. Include nationality and life dates. For unknown creators, use one of the conventions illustrated in the following examples: "unknown," "unknown Chinese," "Chinese," or "unknown 15th-century Chinese."

Tagging examples:

<cdwalite:displayCreator>Michel Erhart (German, ca. 1440-after 1522)</cdwalite:displayCreator>

<cdwalite:displayCreator > probably designed by Giambologna (Flemish, 1529-1608, active in Italy); casting attributed to Pietro Tacca (Italian, 1577-1640) </cdwalite:displayCreator>

<cdwalite:displayCreator > Katsushika Hokusai (Japanese, 1760–1849); published by Nishimura Eijudo (Japanese, 19th century) </cdwalite:displayCreator>

<cdwalite:displayCreator>unknown Chinese</cdwalite:displayCreator>

Display examples:

Creator: Michel Erhart (German, ca. 1440-after 1522)

Creator: probably designed by Giambologna (Flemish, 1529-1608, active in Italy); casting attributed to Pietro Tacca (Italian, 1577-1640)

Appendix 2

Excerpt from CDWA Lite xsd version 0.06 (April 2005)

```
1   <?xml version="1.0" encoding="UTF-8" standalone="yes"?>
2   <!-- edited with XMLSpy v2005 sp1 U (http://www.xmlspy.com) by Ameer Ahmed (Artstor / Switch) -->
3   <!-- This version of XML Document corresponds to CDWA_Lite_Schema_Draft.doc v0.06 changes provided by Getty -->
4   <!--W3C Schema generated by XMLSpy v2005 sp1 U (http://www.xmlspy.com)-->
5   <xs:schema xmlns:cdwalite="http://www.getty.edu/CDWA/CDWALite" xmlns:xs="http://www.w3.org/2001/XMLSchema"
    targetNamespace="http://www.getty.edu/CDWA/CDWALite" elementFormDefault="qualified" attributeFormDefault="unqualified
    ">
6   <xs:element name="AdministrativeMetaData">
7   <xs:complexType>
8   <xs:sequence>
9   <xs:element ref="cdwalite:rights" minOccurs="0" maxOccurs="unbounded"/>
10  <xs:element ref="cdwalite:recordWrap" maxOccurs="unbounded"/>
11  <xs:element ref="cdwalite:resourceWrap" minOccurs="0"/>
12  </xs:sequence>
13  </xs:complexType>
14  </xs:element>
15  <xs:element name="DescriptiveMetaData">
16  <xs:complexType>
17  <xs:sequence>
18  <xs:element ref="cdwalite:objectWorkTypeWrap"/>
19  <xs:element ref="cdwalite:titleWrap"/>
20  <xs:element ref="cdwalite:displayCreator"/>
21  <xs:element ref="cdwalite:indexingCreatorWrap"/>
22  <xs:element ref="cdwalite:displayMeasurements" minOccurs="0"/>
23  <xs:element ref="cdwalite:indexingMeasurementsWrap" minOccurs="0"/>
24  <xs:element ref="cdwalite:displayMaterialsTech"/>
25  <xs:element ref="cdwalite:indexingMaterialsTechWrap" minOccurs="0"/>
26  <xs:element ref="cdwalite:displayStateEditionWrap" minOccurs="0"/>
27  <xs:element ref="cdwalite:styleWrap" minOccurs="0"/>
28  <xs:element ref="cdwalite:cultureWrap" minOccurs="0"/>
29  <xs:element ref="cdwalite:displayCreationDate"/>
30  <xs:element ref="cdwalite:indexingDatesWrap"/>
31  <xs:element ref="cdwalite:locationWrap"/>
32  <xs:element ref="cdwalite:indexingSubjectWrap" minOccurs="0"/>
33  <xs:element ref="cdwalite:classWrap" minOccurs="0"/>
34  <xs:element ref="cdwalite:descriptiveNoteWrap" minOccurs="0"/>
35  <xs:element ref="cdwalite:inscriptionsWrap" minOccurs="0"/>
36  <xs:element ref="cdwalite:relatedWorksWrap" minOccurs="0"/>
37  </xs:sequence>
```

PROGENETOR: An editorial framework for reuse of XML content

*Leen Breure**

The use of structured mark-up, like XML, increases precision in location and retrieval of content and opens up the possibility of reuse. However, if content is simply copied for that purpose, a problem of redundancy will arise. The XML standard itself comprises solutions in the form of advanced hyper linking and virtual documents to prevent redundancy.
This paper discusses a strategy and a tool (Progenetor), to embed such solutions in a normal authoring process in a user-friendly way. Progenetor is *not* another XML editor, but an editorial framework for working with XML document collections. It combines existing XML tools with its own utilities to analyze text, to select, to gather and to rearrange fragments for reuse, and to generate virtual documents. The framework is targeted at publishing through software for dynamic websites, like Cocoon and eXist.

Key words: reuse & repurposing of content, virtual documents, virtual authoring, web publishing, XML, XSLT.

The diversity of the XML editorial environment

An increasingly wide range of XML editors and additional software, both commercial and open source products, fills the current market of electronic publication software. This situation stems from a combination of factors: XML's own versatility, the multitude of vastly disparate application purposes and functionalities, and complex usability requirements.

A run-of-the-mill XML editor may be used in combination with encoding conversion software, an external schema validator, a style sheet designer, an XSLT processor and a repository (XML database) for storing and publishing. The built-in combination of features of a full-fledged editor will guarantee a more stable product than a set of separate tools, but 'one size does not fit all'. How to design an easy to use editor for TEI-based literary text, historical documents, 'classical' structured data, RDF meta-information, and all this preferably in way that conceals the technical intricacies from the user?

It is, therefore, not surprising, that recent reviews of XML editors [3, 4, 15] find considerable differences in features implemented, as with regard to validation against an XML schema, auto completion of elements, the representation of document structure, the support for different schema languages (like RELAX NG and Schematron), XSLT transformations, support of XSL-FO, and different views on the documents being edited (code view, tags-on view, text-only view). When it comes to the management of XML content, most editors offer hardly anything more than a list of documents to be included in a project. There is a wide gap in functionality between this and the far more complex life cycle based XML content management systems and publishing environments, such as LiveLink, Documentum, X-Hive/Docato, DITA (IBM) [5] or Empolis.

*Department of Information and Computing Sciences, University of Utrecht, The Netherlands

A generic editor can be instantiated to yield an editor for a specific document type. The completion of forms has given rise to special XML form editors. IBM offers Xeena, an XML editor, which can be customized for a particular DTD and which allows special icons to be associated with DTD elements. An open extensible framework as Eclipse provides an other solution through editor plug-ins with different functionality. A more fundamental approach is to be found in a new generation of stronger syntax directed editors, in which text editing operations are tighter coupled with document structure [16].

XML content can be published in different ways. Content management systems, text repositories (like Zope), native XML databases (as eXist, X-Hive, Tamino) provide extensive facilities for that purpose. A widely used, dedicated environment for electronic publishing is Cocoon, which allows dynamic transformations of the source content supplied. Content management and electronic publishing are, of course, on the fringe of the editorial environment itself, but an XML editor should at least be able to communicate with these back-end systems.

The Document Life Cycle

Specific editorial requirements depend on the state of the document and the task carried out. A *Document Life Cycle* (DLC) describes how, in general, documents are created, assembled, laid out and published. Recently, Boonstra, Breure and Doorn [1] have proposed a DLC for historical information. The first part is related to the creation process. It comprises the capturing of raw source data, followed by enrichment with meta-data, and further critical editing and reviewing. The second part covers the deployment stage, which consists of retrieval, analysis and presentation (or publishing). In this paper, we shall focus on a more specialized DLC, namely the *Editorial Life Cycle* (ELC), covering the various activities needed for the production of new stories from existing texts (figure 1).

The use of structured mark-up, like XML, has increased precision in information retrieval and has carried the promise of reuse from the beginning. In the historical and cultural heritage domain, a need for reuse of content may arise, when a product has to be tailored to different audiences or platforms (*e.g.* mobile devices), or when a thematic anthology from various sources is compiled, or when dynamic web presentations are generated. It is often motivated by the need for better use of the assets available [13, 11], as Dempsey argues for in his plead for a cultural heritage framework in a shared network space:

> 'A major feature of the new is that fluidity replaces fixity as a dominant characteristic of resource creation and use. Fluid because data flows: it can be shared, reused, analysed; can be adapted, reconfigured, copied, and newly combined in ways which were not possible before. A resource dissolves into multiple individually addressable resources, or can be aggregated in multiple combinations... The creation and use of flows in a digital medium offer unprecedented flexibility, enhancing and augmenting services.' [6]

If the digital resources are regularly replenished and refined, each release may profit from an update of the underlying sources, provided the product is not created as a static text. This process requires not only the use of dynamic documents, but also the application of systematic procedures for extracting, assembling and publishing content components. This, in turn, will make specific demands on the editorial environment. The main question is, how to embed procedures and techniques in the production process, in a way convenient for the author-editor. Note, that the term 'editorial' has got a double meaning in this context: it refers not only to the editor, i.e. 'the software package', but also to 'the human editor' and its editorial work (in Dutch: *redacteur* and *redigeren*).

A new document could be simply produced by copying fragments from existing sources, as we did in the pre-computer era. This, however, has as major drawback the creation of redundancy. This problem is well known in database design: if the same data are stored on different places, it is more difficult to guarantee consistency when the information is regularly updated. If we imagine a repository of digital content, the redundancy may become easily unmanageable likewise. The XML standard offers a solution in the form of advanced methods of hyper linking (XLink) and content including (XInclude), thus creating virtual documents, generated on demand. Like views on a database, they are instantiated when required. Ideally, these documents do not contain much data of their own, and consist of references to underlying digital sources only.

There is another problem with simple copying. The set of copied fragments itself will rarely result in an

acceptable story, as we all may have experienced when we tried to insert apparently ready-made sentences from previous publications into a new one and ended up rewriting a considerable part. Some passages may resist reuse by phrasing. Back references at the beginning of a paragraph in the form of 'this person', 'he', 'this event', 'when this happened', etc. may obstruct the isolation of such a text fragment as a reusable component. The source text has to be modified, some linking text is to be inserted in the target document, or we must provide a modified version. This complicates the reuse of narrative text, in comparison with database data.

Therefore, reusability has proven to be most successful with technical documentation with a high degree of control over authoring. Self-contained components are a key prerequisite: each unit must stand on its own and not depend on other parts of the documentation to render it useful. In addition, authors should strictly adhere to style guides to produce uniform texts [2]. Unfortunately, the creation of the majority of valuable content is beyond such strict control as in the publication of technical manuals. At the same time, some form of reuse may be still worth considering, because the production process is labor-intensive [9, 10, 14].

The solution described here comes close to the usual authoring process with copy-and-paste, but avoids the problems mentioned by preceding content analysis and by removal of redundancy during post processing. The process can be represented by means of an ELC (inside a more general DLC – figure 1) and comprises the following stages:

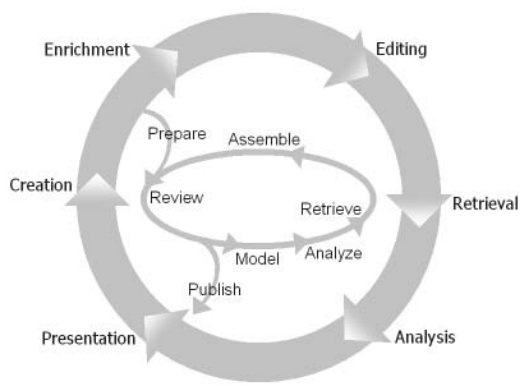


Figure 1. The editorial life cycle for reuse of content

1. **Preparing**. Although we use existing digital material, this may not always be in the right format. XML processors are quite picky with regard to encoding issues. An XML document must not only be well formed, but also fully match the encoding standard chosen (*e.g.* single byte ISO-8859-1, or multi-byte Unicode: UTF-8, UTF-16).
2. **Reviewing**. Reuse of content will start with a careful study of the available documents. In the beginning, this may be a matter of simple reading in order to gather basic information for more advanced explorations. We tacitly assume a leading question, which will guide further activities and which will provide a purpose to the following content assembling. The main deliverable of this stage is the description of the initial interest as a subject that can be actually traced in the material available, and some idea how this could be made into the new product envisaged.
3. **Modeling**. The content as it exists may not yet have been properly modeled in the form of a DTD or XML schema. In addition, an existing model should also be evaluated in view of the product to be generated, which may require a different structure. Such a comparison provides a first estimate of the complexity of the intended reuse.
4. **Analysis**. As a sequel to reviewing and modeling, content may be explored in a more sophisticated way, by means of queries, concordances, or statistical text analysis (word frequencies, collocation analysis, etc.). A qualitative study will lead to a semantic field: a set of key words describing the subject. It may be helpful to edit the source text and to encode the semantics by adding additional attributes to the related elements in the source (enrichment). A quantitative analysis produces measures for text length. If applied to both source text and texts of the target genre, it will provide an exact insight in the effort required for reuse. How much text is to be summarized, and what must be elaborated? Is the subject described through discourse elements that fit the type of text envisaged? In combination with document modeling these data answer the question what the content transformation will cost.
5. **Retrieval**. Now we are ready to collect the content components for the new story. An XML query facility is needed that allows evaluation and selection of discrete elements, and delivers the results in a predefined document structure suitable to subsequent assembling and publishing.

6. **Assembling**. We assume a kind of text compilation that requires the explicit intervention of a human author-editor. In this stage, XML elements may be rearranged, and, with some reluctance, modified if reuse as such is impossible. New linking text can be added where required.
7. **Reviewing**. Once assembled, the new content has to be reviewed and evaluated for compliance with the purpose, subject and genre chosen.
8. **Publishing**. Finally, the content is to be marshaled for publishing. The assembled text is automatically compared with its underlying sources, and converted to a virtual document in which redundancy has been removed. Some other conversions may be needed to match the technical requirements of the publishing platform.

Progenetor as an editorial framework

*Architecture and Processing Model***.** Progenetor is a framework for XML tools. It comprises template-based utilities, defined by means of special tags to control processing, to execute XSLT code and to call other text software. Thus it provides a modular service layer, which facilitates tasks concerning the preparation of XML text, the disclosure of XML documents, the retrieval of text fragments, and the creation of virtual documents.

In comparison with a framework as Eclipse, Progenetor offers only a loose integration, without relying on an API. The programs invoked are mostly console applications, which run without a graphical interface and get their input through command line parameters. Progenetor presents a uniform user interface for input of instructions and for display of generated output, offering a substitution for scripts and batch files. Output produced by other programs can be captured and displayed through Progenetor's own user interface (without opening a DOS window).

Progenetor is a hypermedia application and may be best conceived as the hybrid merger of a HTTP client, a server back-end and an operating system shell program. It is a single user, stand-alone program for the Windows platform, written in Delphi, with a built-in web browser component (fully Internet Explorer compatible), which allows a highly customizable HTML user interface. The XML utilities are based on an underlying XSLT library, which is accessible to users and may be modified as desired. The associated external programs comprise well-known XML tools, editors with special features, and corpus linguistic software. Utilities can be chained together, yielding wizard-like editorial pipelines, which simplifies the application of procedures to a specific set of documents. The XML input may vary from a single document to a collection. Collections are defined as a simple list of files, or through a directory path with wild cards.

Operational tasks are grouped in utilities. A utility consists typically of:

1. a HTML *form* for specifying the task parameters (like filenames and processing details), and
2. a *template* containing some of Progenetor's dynamic tags. A dynamic tag has as general format: <*#tag-name parameter1*='...', *parameter2*='...' etc.>.

Templates and auxiliary programs are referenced by means of *symbolic names* to make utilities independent of a specific computer configuration. Progenetor uses its own registry for mapping these symbolic names to physical file names.

When a form is submitted, the program reads the form data, and checks, whether the URL may point to a web site somewhere, or refers to a Progenetor job. In the latter case the form must reference a specific template to be instantiated. The output generator will replace the template's tags with data produced by the dynamic tags. In this way, a new document is created, which can be saved, displayed or, in case of a script, executed. A template may model any kind of text, not necessarily an XML document. With equal ease Progenetor will instantiate a HTML file, any scripting language program, an XSLT file, or a Windows batch file. If required, it can execute the code generated in the same process.

The Progenetor template is a versatile concept: it can be a 'template for a document' or a 'template for a process', and both aspects may be mixed. Some tags are replaced with text (*e.g.* output from an XSLT processor), while others are used only to instruct the processing environment, yielding no replace text at all. This category can preprocess a text, manipulate variables, tell Progenetor in which frame output is to be displayed, or start an other program. When a template contains mainly the latter type of tags, it comes close to a 'script'. However, the more than one hundred dynamic tags work well as a template language (i.e. to form a list of statements), but should not be considered as a scripting language. Conditional evaluation of template sections, string manipulation and data per-

sistence through stored variables are all provided, but the mechanism of dynamic tags don't measure up to what one may expect from a real scripting language with control structures, data typing, operators and expressions. For such programming purposes JavaScript can be used very well in the HTML form associated with the template: the results can be stored in fields and sent to Progenetor's back-end on submit.

Reusability Support. Progenetor comes with a set of utilities, which serves the double purpose of useful tools and examples of how editorial support for content reuse may be implemented. Because of the HTML interface and the modular framework, users can easily change this default set. The current version supports the ELC outlined above through:

– management of XML document collections, stored in the local file system
– encoding conversion
– automatic ID generation
– creating a table of contents
– XML modeling (DTD, W3C Schema, RELAX NG, tree diagrams)
– integrating external search-replace and concordance software
– text exploration and analysis: a form interface on top of XSLT library routines for XPath queries and simple text statistics.
– document assembly for content reuse: a 'grabber' for collecting fragments from source documents, options for re-editing and rearranging them, conversion to non-redundant skeletons (virtual documents), or to XInclude files.
– reviewing final products
– accessing an XML repository, such as a native XML database.

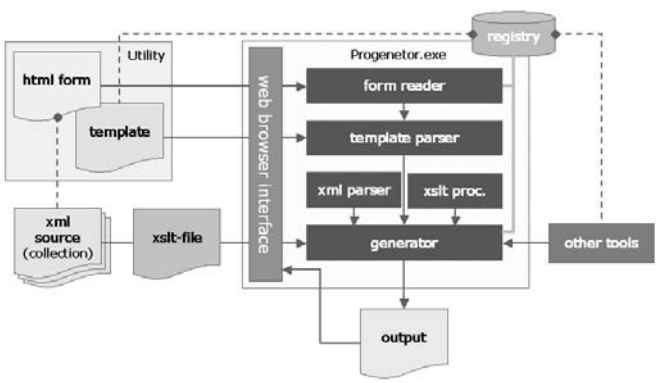


Figure 2. The Progenetor architecture

Reusing Content. Progenetor's content assembling capabilities may be best illustrated by means of one of the small projects, in which the system was tested. It is aimed at building a new biography of the Austrian painter Gustav Klimt (1862-1918) out of different Internet resources, in principle without modifying the original text fragments, so, by actually reusing content components where possible.

The Web sources were downloaded as HTML documents and first stripped of all tags. Next, the plain text was encoded in XML, using a DTD that emphasized the events, oeuvre, style, artistic environment, contemporary reactions, and the current art historian's assessment. Some light pre-editing was required to correct back references, and to convert a few unsuitable passages into comments.

Progenetor utilities were used in all subsequent steps. The texts were listed in a single collection file in order to process them together. A unique ID-attribute was automatically added to all major XML elements. Now, the text could be explored by entering element names and key words, using a utility, which builds and evaluates complex XPath expressions backstage. The results were further refined by calculating simple text statistics for subsets matching a specific XPath expression, thus estimating the homogeneity of the material ('how many tags X are in each document and how much text do they enclose?')

This exploration gave a good impression of new story lines feasible and the appropriate query formulations. It was decided to regroup the material into a few short chapters: profile, style, oeuvre, contemporary reactions, life in Vienna, and chronology. With the help of the grabber utility fragments were retrieved in a list with check boxes, allowing the author-editor to include and exclude distinct elements as desired (figure 3). This produced a raw, but well formed XML document, which was saved and further edited for the assembling stage. Finally, Progenetor was used to reduce the edited text to a skeleton. In this operation new and modified elements are preserved in full, while unmodified elements are replaced with references to the respective

source documents, thus removing redundancy. This process is reversible without any loss of information: a skeleton can be easily converted back to a complete XML document, which may be re-edited. Alternatively, Progenetor can generate an XInclude file.

Conclusion and future work

Editorial functionality is currently prepacked in ready-made software, leaving only limited room for customization, while editorial tasks are growing more complex, particularly when content is to be reused in different contexts and for different target groups. Progenetor has been developed to experiment with reusing content. In addition it is proof of concept of integrating a variety of tools in a flexible framework on basis of command line calls. The concept of a modular framework is not new. XML editor plug-ins are now available for Eclipse, and Cocoon follows a modular approach for publishing. However, these systems do not (yet) cover the specific editorial functionality required here.

The reusability technique discussed above bears some resemblance to reuse by copy-and-paste, but distinguishes itself by systematic modeling and analysis of both source and target genres, and by control of redundancy. With regard to methodology, there remains much to do:

– An experiment as that with Klimt's biographies is only intended as an illustration of the system's technical capabilities. It does not say much about what makes content reusable. More research is required to distill an empirical theory, which helps to predict the efforts required for reuse of content between one genre and an other [2]. Or, the other way around, to specify in advance the features and constraints of reusable content to be created.
– A document model is often considered as a static set of rules controlling an editorial project. The reuse of content asks for greater flexibility and for techniques that makes it easy to redefine XML content on basis of a different DTD without extensive manual recoding.
– In addition, we need better tools to model the structure of document *instances*. XML modeling is usually directed to document classes. However, when content is reused, the new variants generated may share the same DTD, while the stories differ in a significant way, which requires documentation as well. However, adequate modeling techniques on this level are still lacking. Rhetorical Structure Theory (and its accompanying diagram technique) [12], fre-

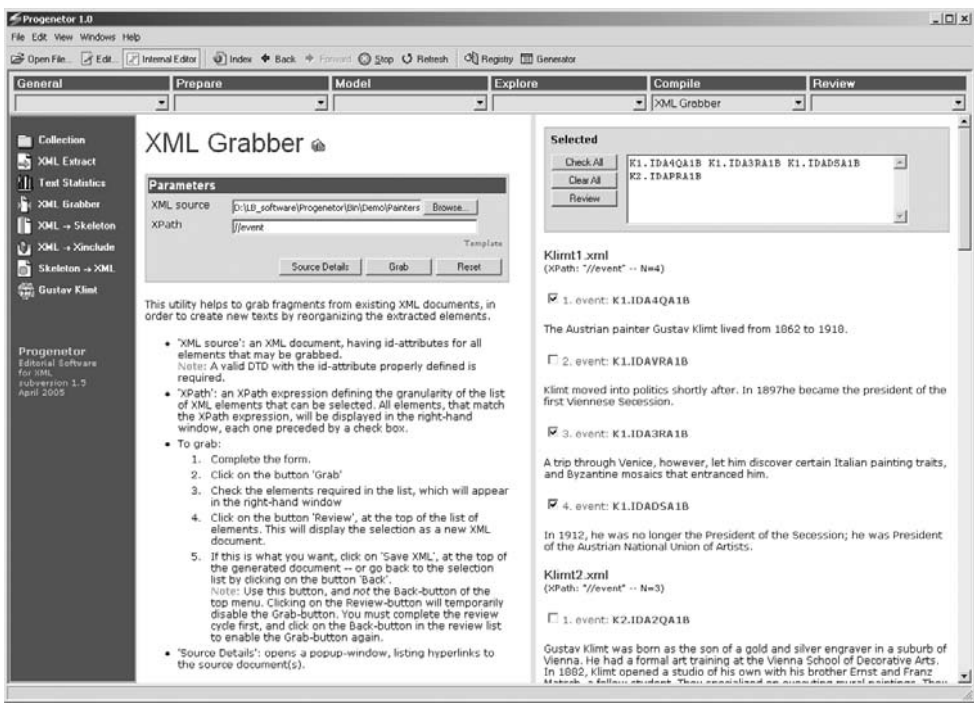


Figure 3. Progenetor's grabber interface to retrieve content components for reuse

quently used to model a specific discourse structure, is not suitable for these more technical purposes.

These desirables outline some of the future work in this project, which will move more into a methodological direction, rather than develop its own technical infrastructure.

References

[1] Boonstra, O., Breure, L., & Doorn, P. (2004). *Past, present and future of historical information science.* Amsterdam [NIWI-KNAW]: http://www.cs.uu.nl/research/projects/i-cult/Publications.htm

[2] Breure, L. (2005). Reuse of content and digital genres. Oostendorp, H. van, Breure, L. & Dillon, A. (eds.). Creation, use and deployment of digital information [Lawrence Erlbaum Associates]. Mahweh (New Jersey), London.

[3] Broek, T. van den (2004a). *Backing the right horse. Benchmarking XML editors for text-encoding.* Utrecht: http://www.cs.uu.nl/research/projects/i-cult/Publications.htm

[4] Broek, T. van den (2004b). *XML Editors: Which one to choose?* Information paper [Oxford Text Archive]. Oxford, 31 October 2004: http://www.cs.uu.nl/research/projects/i-cult/Publications.htm

[5] Day, D.R., Priestly, M., & Schell, D.A. (2003). Introduction to the Darwin Information Typing Architecture. Toward portable technical information: http://www-106.ibm.com/developerworks/xml/library/x-dita1/index.html

[6] Dempsey, L. (2000). Scientific, industrial, and cultural heritage: a shared approach. A research framework for digital libraries, museums and archives, in: *Ariadne*, issue. 22, 12 January 2000: http://www.ariadne.ac.uk/issue22/dempsey/

[7] Eclipse Platform Technical Overview [Object Technology International, Inc. February 2003]: http://www.eclipse.org/articles/index.html

[8] Falkovych, K., Nack, F., Ossenbruggen, J. van & Rutledge, L. (2003). *Semantics in multi-facet hypermedia authoring* [Centrum voor Wiskunde en Informatica: report INS-E0307].

[9] Garzotto, F., Mainetti, L., & Paolini, P. (1996). Information reuse in hypermedia applications. *Hypertext* 96, 93-104.

[10] Glushko, R., & McGrath, T. (2002). Patterns and reuse in document engineering: http://www.idealliance.org/papers/xml02/dx_xml02/papers/03-04-01/03-04-01.pdf

[11] Kansa, E. & Schultz, J. (2004). *Perspectives on cultural heritage and intellectual property* [Alexandria Archive Institute]: http://www.alexandriaarchive.org/AAI%20IP%20Whitepaper.pdf

[12] Mann, B. (1999). *An Introduction to Rhetorical Structure Theory (RST)*: http://www.sil.org/~mannb/rst/rintro99.htm

[13] McKenna , G. (2004). *Framing the reuse of digital cultural heritage* [European Museums' Information Institute]: http://www.innovations-report.com/html/reports/information_technology/report-31328.html

[14] Nanard, M., Nanard, J., & Kahn, P. (1998). Pushing reuse in hypermedia design: golden rules, design patterns and constructive templates. *Proceedings of the ninth ACM Conference on Hypertext and Hypermedia: links, objects, time and space-structure in hypermedia systems*, 11-20, http://doi.acm.org/10.1145/276627.276629.

[15] Paul, R. (2005). Open source XML editors examined, in: *News Forge*. The online newspaper for Linux and Open Source, Monday February 28, 2005: http://programming.newsforge.com/article.pl?sid=05/02/24/1650248

[16] Schrage, M.M. (2004). *Proxima. A presentation-oriented editor for structured documents*. PhD thesis [Instituut voor Informatica en Informatiekunde, Universiteit Utrecht].

The early letters of The Royal Society 1657-1741: Managing diversity and complexity

*Jan Broadway**

Background

The letterbooks of the Royal Society[2] represent an important source for the study of scientific endeavour in the late seventeenth and early eighteenth centuries. They include letters received by the society or its members, copies of outgoing correspondence and various miscellaneous documents. There are accounts of experiments and observations, often with accompanying illustrations. There are letters relating to the formal business of the society and its affiliated societies, from provincial gentlemen seeking to establish their scientific credentials, and from members of the society on continental tours. There is correspondence with scholars in the Netherlands, France, Germany and elsewhere, much of it preserved in the original language and in translation. The letterbooks do not represent a complete archive of the correspondence of the society. There are gaps and omissions, as the maintenance of the records relied upon the enthusiasm and diligence of successive secretaries. As such the collection mirrors the varying levels of activity within the society over the first decades of its existence: decades of alternating enthusiasm and despondency, activity and desuetude.

The letterbooks as currently bound are arranged alphabetically by letter writer and chronologically within that ordering. This arrangement makes it straightforward to examine the letters written by individuals, but locating the letters they received is more difficult. Researching networks of correspondence or themes within the archive requires more effort still. In the past the collection has predominantly been mined by historians of science interested in particular scientists. Many of the letters associated with important individuals such as Newton, Boyle and Oldenburg have appeared in edited collections, but much of the archive remains relatively unexplored. The letters are currently being catalogued in a project financed by the Andrew Mellon Foundation. However, it is difficult to do justice to the diversity of subject matter and the intricate interconnections between many of the letters within the confines of a formal catalogue.

The project

In 2003 CELL became involved in discussions with the Royal Society and the Newton Project[3] about making the letterbooks available on-line. It was appreciated that the aims of the parties were different and that the project would need to accommodate these. The archivists were predominantly interested in making the collection accessible to a wider readership and in providing images that would reduce the need for researchers to access the fragile originals. The academics were interested in exploring the complexity of the resource, enabling new avenues of research to be pursued and relating the contents of the letterbooks to other sources, such as the society's minutes and publications. The personnel at CELL have experience in reconciling the different aspirations of academics and archivists and it was agreed that it would be appropriate for CELL to take the lead in this project.[4] Although considerable support was expressed by researchers in the field, two applications for funding were unsuccessful. In November 2004 Professor Lisa Jardine, the director of CELL, led a public masterclass on 'Exploring the Early Letters of the Royal Society'. Following this it was agreed that a pilot project would be undertaken, which would provide a basis for realistic project plan-

* AHRC Centre for Editing Lives and Letters, Queen Mary, University of London[1]

ning and a working example of what we were trying to do. The pilot project has been funded out of CELL's core funding.

As technical director of CELL, I was responsible for the pilot. I am an early modern historian with experience in XML encoding and general purpose programming. The pilot project took 12 letters for 6 of which we had images and produced an edition presented through a web site: http://www.livesandletters.ac.uk/rs/pilot/main.html. This took approximately 20 project days over 5 months to complete and was launched on 25th April 2005, to coincide with the rerunning of Professor Jardine's masterclass.

Structure of the edition

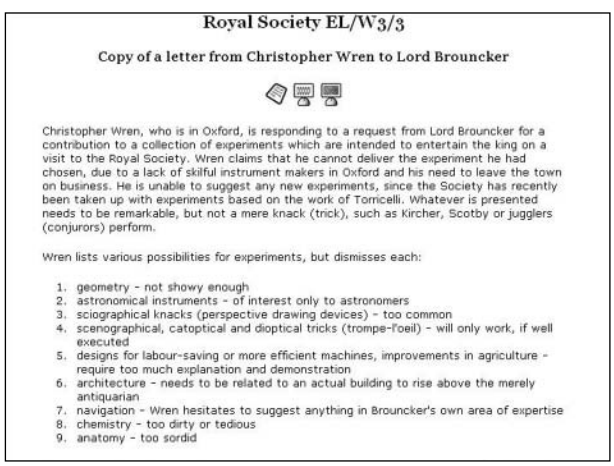
It was agreed that the initial aims of the Early Letters project would be to provide on-line for each letter within the archive:

– a discursive description
– images of the original
– a transcription

There would also be mechanisms by which readers could negotiate the collection in various ways, allowing them to pursue themes, reconstruct correspondences, etc. The second stage would link the letterbooks to the society's minutes, the *Philosophical Transactions* and manuscript sources in the Royal Society's archives and elsewhere.

The pilot project involved:

– creating a transcription policy;
– creating an XML encoding policy for the transcriptions and supporting documents;
– transcribing the sample letters;
– preparing the images;


Royal Society EL/W3/3

Copy of a letter from Christopher Wren to Lord Brouncker

Christopher Wren, who is in Oxford, is responding to a request from Lord Brouncker for a contribution to a collection of experiments which are intended to entertain the king on a visit to the Royal Society. Wren claims that he cannot deliver the experiment he had chosen, due to a lack of skilful instrument makers in Oxford and his need to leave the town on business. He is unable to suggest any new experiments, since the Society has recently been taken up with experiments based on the work of Torricelli. Whatever is presented needs to be remarkable, but not a mere knack (trick), such as Kircher, Scotby or jugglers (conjurors) perform.

Wren lists various possibilities for experiments, but dismisses each:

1. geometry - not showy enough
2. astronomical instruments - of interest only to astronomers
3. sciographical knacks (perspective drawing devices) - too common
4. scenographical, catoptical and dioptical tricks (trompe-l'oeil) - will only work, if well executed
5. designs for labour-saving or more efficient machines, improvements in agriculture - require too much explanation and demonstration
6. architecture - needs to be related to an actual building to rise above the merely antiquarian
7. navigation - Wren hesitates to suggest anything in Brouncker's own area of expertise
8. chemistry - too dirty or tedious
9. anatomy - too sordid

Figure 1. Part of the description for EL/W3/3

– writing an editorial framework;
– providing biographical references;
– writing code to transform the XML into a web site.

Since we do not expect to be able to fund the whole project from a single source and anticipate that descriptions, transcriptions and images will be added incrementally to the web site over time, we wanted to develop a methodology and tools which would support this. Because of the eclectic nature of the collection, which includes many copies and extracts rather than original letters, we decided that the detailed manuscript transcription policies adopted by the Newton Project or the Boyle Workdiaries[5] would be inappropriate. We also felt that an accessible transcription would be more in keeping with the aims of the project. Our purpose is to create texts than are accessible to the interested reader, not to painstakingly mirror all the features of the manuscripts and to this end we adopt the following conventions for the transcriptions:

– The ampersand (&) is retained.
– Thorn is silently altered to 'th'.
– The use of i/j and u/v is modernised.
– The long s is silently altered to 's'.
– Contractions are silently expanded.
– Alchemical symbols are interpreted as their planetary or chemical equivalent, which ever makes most sense in the context.

Apart from the above conventions, original spelling and punctuation are retained. Insertions and corrections are silently incorporated into the text, while deletions are ignored.

The transcription itself is presented to the reader without editorial intervention. Each letter is provided with a discursive description, which is used to identify people and subjects referred to in the text and enables the editor to raise any issues or points of interest.

This description for EL/W3/3 there are links to the biographical register for Christopher Wren, Lord Brouncker, Torricelli, Kircher and Scotby. Thematic links were not implemented in the pilot, but this extract provides a good illustration of the range of subjects than can be covered by a single letter. It also illustrates how scientific language has altered – sciographical, scenographical, catoptical, etc. are not in common use today. The advantage of the discursive introduction is that such terms can be explained or given their modern equivalents, without overburdening the transcription with mouseovers and links.

This discursive approach is far more flexible than a traditional catalogue, as can be seen by taking the existing entry for EL/W3/3 from the Royal Society's online catalogue:

Repository	GB 117 The Royal Society
Level	Item
RefNo	EL/W3/3
Title	Copy of a letter from Christopher Wren to William Brouncker
Date	30 July 1663
Description	Containing suggestions, on scientific themes, for Charles II's entertainment at a reception
Extent	3 sides
Format	Manuscript document
Language	English

The descriptions are also used to link letters together. For example, in one of our sample letters some information is requested, the receipt of which is acknowledged in another. Currently such links help the reader to follow a correspondence, but in the future they could be used to generate semi-automatic thematic indexes of related letters.

As currently implemented the interface to the letters is through the description, which provides access to the transcription and, if images are available, to the images and to the transcription and images in parallel. If viewed on its own the transcription is presented as a single document. When viewed in parallel, the transcription is divided into pages to match the images. The purpose of the descriptions is to aid the non-expert reader to understand the documents, while providing the editor with freedom to explore the content discursively and to create cross-references without the restraints that are often imposed by utilising a transcription for this purpose. It is envisaged that in the future different groups of readers (secondary level students, undergraduates, academics, the general public) would be presented with different mediations to the material. One area we want to explore is how these different groups react to the descriptions and how they utilise them. We suspect that some academic researchers will prefer unmediated access to the images, which is not currently accommodated.

Implementation

All the content for the pilot is encoded in XML: transcriptions, descriptions and supporting material. The use of TEI[6] was considered, but rejected in favour of a set of elements and attributes that was designed specifically for this project. The set was kept deliberately simple, as we want to avoid complexity in the encoding and concentrate on the substantive content of the documents.

For each letter there is a single XML encoded file, which contains both the description and the transcription. The XML and image files are named according to their archival reference. For example, Letterbook A item 31 has three pages, while letterbook B1 item 61 has only 1.

EL.A.31.xml	EL.B1.61.xml
EL.A.31 1.jpg	EL.B1.61.jpg
EL.A.31 2.jpg	
EL.A.31 3.jpg	

The web site is created by processing the content offline using a program written in Python that utilises the DOM interface. Python was chosen because of its suitability for rapid prototyping and its support for object orientation. The programme has been written to allow either the generation of the entire web site or selective update of specified files, permitting incremental addition of content. The programme makes several passes through the XML, in order to generate the content and infrastructure of the web site.

Early Letters DTD:

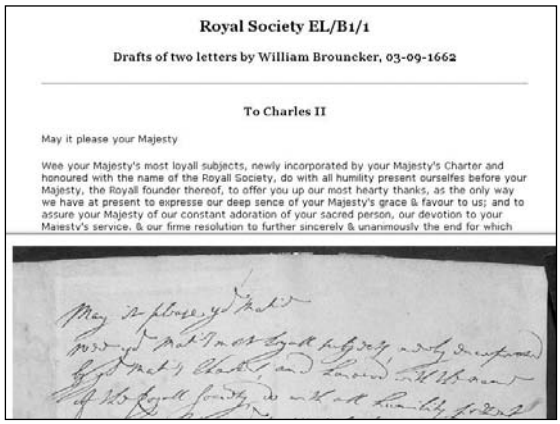
```
<!ELEMENT letter_xml (title, order_date, description, section+)>
<!ELEMENT title #PCDATA>
<!ELEMENT order_date EMPTY>
<!ELEMENT description (p|list|linkref)+>
<!ELEMENT section (title?, notes?, body, notes?)>
<!ELEMENT notes (note)+>
<!ELEMENT note (#PCDATA)>
<!ELEMENT body (subsection|p)+>
<!ELEMENT subsection p+>
<!ELEMENT p (#PCDATA|person|cite|pb)+>
<!ELEMENT person #PCDATA>
<!ELEMENT cite #PCDATA>
<!ELEMENT linkref (#PCDATA|person)+>
<!ELEMENT list (item)+>
<!ELEMENT item #PCDATA>
<!ELEMENT pb EMPTY>
<!ATTLIST letter_xml archive CDATA #REQUIRED>
<!ATTLIST letter_xml item CDATA #REQUIRED>
<!ATTLIST letter_xml version CDATA #REQUIRED>
<!ATTLIST order_date value CDATA #REQUIRED>
<!ATTLIST section n CDATA #REQUIRED>
<!ATTLIST section type CDATA #REQUIRED>
<!ATTLIST subsection n CDATA #REQUIRED>
<!ATTLIST p align #IMPLIED>
<!ATTLIST person id #REQUIRED>
<!ATTLIST linkref n #IMPLIED>
<!ATTLIST linkref fname #IMPLIED>
```

Some of the letters used in the pilot have a complex structure, so the transcription and description are encoded into matching sections and subsections, which allows direct links to the appropriate part of a long document to be automatically generated by the Python programme. EL/B1/1 includes two separate texts in a single archival reference and is encoded as two sections. EL/A/28, EL/A/29 and EL/A/30 include copies of the minutes of the Dublin Philosophical Society with a covering letter. They are encoded as two sections for the minutes and the letter, with the minutes subdivided into further subsections for the minutes of each separate meeting. Links to descriptions for other archival references are added by the editor using the linkref element.

In order to create the web site the programme:

1. generates a HTML document containing a transcription for each letter;
2. if there are images for the letter:
   a. if there is more than one page, generates a HTML document for each page incorporating appropriate links for next and previous;
   b. generates a HTML document to allow access to the images;
   c. generates HTML framesets to enable access to the image and corresponding transcription for each page that includes the start of a section;
3. generates a HTML document containing a description for each letter and links to the transcriptions and images as appropriate;
4. generates an index for all letters ordered by date (using the value attribute of the order_date element);
5. generates an index for all letters ordered alphabetically by archival reference (using the archive and item attributes of the letter_xml element);
6. generates the editorial content for the site;
7. generates the biographical index.

If there are command line arguments specifying letter identifiers, only those letters are processed in steps 1, 2 and 3, followed by the regeneration of the indexes


Royal Society EL/B1/1

Drafts of two letters by William Brouncker, 03-09-1662

To Charles II

May it please your Majesty

Wee your Majesty's most loyall subjects, newly incorporated by your Majesty's Charter and honoured with the name of the Royall Society, do with all humility present ourselfes before your Majesty, the Royall founder thereof, to offer you up our most hearty thanks, as the only way we have at present to expresse our deep sence of your Majesty's grace & favour to us; and to assure your Majesty of our constant adoration of your sacred person, our devotion to your Maiestv's service. & our firme resolution to further sincerelv & unanimouslv the end for which

Figure 2 The parallel transcription and image for EL/B1/1

in steps 4 and 5. Groups of letters can be specified by using wildcards *e.g.* 'EL.A.*' for all letters from letter-book A. If the command line arguments specify files containing editorial content or the biographical index, either step 6 or 7 is run as appropriate.

Conclusions

Although the selection of letters in the pilot was small, it has enabled us to better understand the complexity of the material. It has convinced us that the needs of the archivists and researchers will be best met by keeping the transcriptions simple and concentrating the bulk of our efforts on the supporting scholarly apparatus. This selection of just 12 letters has produced a biographical index of 64 individuals. For the pilot the links from the biographical index to the letter descriptions were hard-coded, but we would like in future to generate these links programmatically during the generation of the web site.[7] A rudimentary analysis of the topics covered by the letters in the pilot has confirmed the diversity and complexity of the subject matter of the collection.

The complexity of the topic map produced from this small sample confirmed that a conventional approach to the provision of metadata for the corpus will be insufficient to meet our needs. The next stage of the project is to begin the development of a thesaurus, which will allow us to map the diverse subject matter of the corpus to enable conceptual searching through the Semantic Web. It is intended that this work will take place over the next 4 to 6 months. The inclusion of additional content into the edition will be part of an ongoing process.

Notes

1 Known as CELL, http://www.livesandletters.ac.uk
2. http://www.royalsociety.ac.uk
3. http://www.newtonproject.ic.uk
4. A. Wiggins, 'The Auchinleck Manuscript Project as an exemplar of collaborative research', CELL ref. FOR/2004/04/001,http://www.livesandletters.ac.uk/research/indexpapers.html
5 http://www.livesandletters.ac.uk/wd
6. http://www.tei-c.org – similarities between the TEI Guidelines and the tag set used here are inevitable.
7 This could be accommodated with minor changes to the Python code.

Figure 3. Extract from the Early Letters topic map:

- intellectual societies
  - exchanges between
  - members
  - officials
  - meetings
  - publications
  - facilities / equipment
- scientific phenomena
  - weather
    - observations
    - extreme
    - rain
      - flooding
    - winds
      - trade winds
      - relationship to air pressure
  - physiology
    - human
      - disability
        - squint
      - disease
        - rabies
        - horned girl ?
      - fasting
    - animal
      - dogs
        - experiments
        - biting humans
        - mad
      - fish
- news
  - political events
    - Ireland

Content-Based Art Retrieval (C-BAR)

Egon van den Broek[a,b], Thijs Kok[c], Eduard Hoenkamp[b], Theo Schouten[c], Peter Petieta & Louis Vuurpijl[b]

An online Content-Based Art Retrieval (C-BAR) system is introduced that provides entrance to the digitized collection of the National Gallery of the Netherlands (the Rijksmuseum). The current retrieval system of the Rijksmuseum is textbased and requires expert knowledge concerning the work searched for, else it fails in retrieving it. C-BAR extends this system with querying by an example image, which can be provided to the system or can be selected through browsing the collection. The color and texture features of the example image are extracted and compared with those of the images in the collection. Hence, based on text or on content-based features, the collection can be queried. Moreover, the matching process of C-BAR can be inspected. With the latter feature, C-BAR not only integrates the means to inspect collections by both experts and laypersons in one system but also provides the means to let the user to understand its working. These characteristics make C-BAR a unique system to access, enhance, and retrieve the knowledge available in digitized art collections.

Keywords: Content-Based Image Retrieval, Information Retrieval, C-BAR, color, texture, art

1 Introduction

Vast amounts of digitized online image archives come available (*e.g.*, photo databases and museum collections). In the cultural domain, museums like the National Gallery of the Netherlands (the Rijksmuseum),[1] the Virtual Catalog for Art History (see Figure 1), and the Hermitage museum[2] are extending their reach by making part of their collection available via Internet.[3] These examples are followed by many libraries, museums, and governmental institutes, with the goal to preserve our cultural heritage.

Making these collections publicly available also requires information systems for indexing, browsing, and retrieving the information. However, most of these systems require expert knowledge to use them efficiently. Hence, laypersons will not use these systems and, consequently, will not be able to inspect the collections in a satisfying manner.

This paper describes a Content-Based Art Retrieval (C-BAR) system that provides entrance to the collection of the Rijksmuseum. It provides suitable access to the database of art materials for both experts and

a Department of Artificial Intelligence, Vrije Universiteit Amsterdam, De Boelelaan, 1081a, 1081 HV Amsterdam, The Netherlands {egon,pjpetiet}@few.vu.nl http://www.few.vu.nl/~{egon,pjpetiet}/
b Nijmegen Institute for Cognition and Information, Postbox 9104, 6500 HE Nijmegen, The Netherlands hoenkamp@acm.org l.vuurpijl@nici.ru.nl http://hwr.nici.kun.nl/~vuurpijl/
c Institute for Computing and Information Science, Radboud University Nijmegen, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands T.Kok@student.ru.nl T.Schouten@cs.ru.nl http://eidetic.ai.ru.nl/thijs/ http://www.zs.ru.nl/~ths/

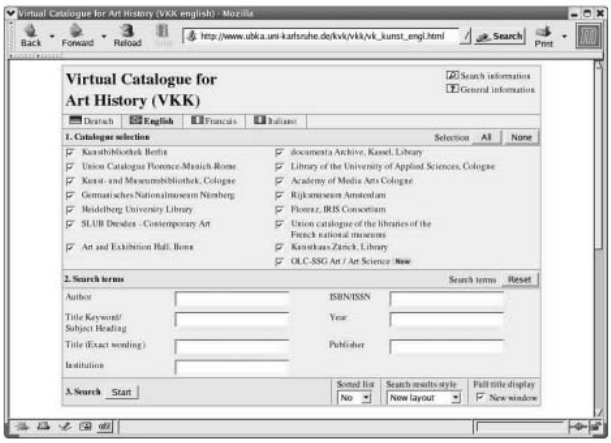


Figure 1. A traditional text-based information retrieval system

laypersons. We start with a brief introduction and comparison of image retrieval by text and by content followed by the introduction of the domain of application, in Section 2.1. The presentation of the results, an often ignored issue, is briefly denoted in Section 3. In Section 4, the online C-BAR system is introduced. We end this paper with a discussion on issues to be solved and topics of future research.

2 Image retrieval: text-based versus content-based

In 1992, Kato[4] introduced the term Content-Based Image Retrieval (CBIR), to describe his experiments on automatic retrieval of images from a database by color and shape features. Since then, CBIR has developed into a separate field.

CBIR is the application of computer vision to the image retrieval problem; i.e., the problem of searching for images in large image databases. Most image retrieval engines on the world wide web (WWW) make use of text-based image retrieval, in which images are retrieved based on their captions, descriptions, and surrounding text. Although text-based image retrieval is fairly successful, it fully depends on the verbal annotations that accompany the images. Consequently, it requires every image in the database or on the WWW to be annotated.

A decade ago, Gudivada and Raghavan[5] identified twelve fields of application in which CBIR can prove its usefulness: crime prevention, the military, intellectual property, architectural and engineering design, fashion and interior design, journalism and advertising, medical diagnosis, geographical information and remote sensing systems, cultural heritage, education and training, home entertainment, and WWW searching. Despite this range of applications, Smeulders, Worring, Santini, Gupta, and Jain6 noted in 2000 that 'CBIR is at the end of its early years' and is certainly not the answer to all problems. Vuurpijl, Schomaker, and Van den Broek[3] concurred with Smeulders et al. with judging that CBIR systems 'are far from mature'. They continue with an explanation, which contains a quartet of arguments: (i) CBIR techniques still yield unacceptable retrieval results, (ii) they are restricted to the domain that is covered, (iii) they lack a suitable user-interface, and (iv) are mainly technology-driven and, consequently, require the use of domain knowledge to fulfill the user's information need. [7, 8]

Vuurpijl et al.[3] also describe a change in research perspective regarding CBIR systems, from computer vision and pattern recognition to cognitive science and psychology. For example, Rui, Huang, and Chang[7] and Jörgensen[9] emphasize how important it is to consider the human in the loop. Using knowledge about the user will provide insight in how the user-interface should be designed, how retrieval results may be presented, and what will categorize the typical information need of the general public.

2.1 Domain of application

Historical archives are rich sources of information, more and more of which available in digitized form. For example, the institute for Dutch history (ING)[10] recently finished a project with a time span of three decades. It recently introduced an online database of all correspondence of Willem of Orange. However, this project merely applies traditional information processing techniques. This paper uses the database of the National Gallery of the Netherlands (the Rijksmuseum). This database is already annotated. These annotations are already used in an online search system.[1]

The Dutch Rijksmuseum states: 'Research is the basic premise of all museum activity, from acquisition, conservation and restoration, to publication, education and presentation. In general, this research involves the object or work of art in the museum as being a source of information.'[1] However, how can these sources of information be efficiently accessed and enhanced and how can knowledge be retrieved from them? The Rijksmuseum has made their collection available through a web-interface.[1] Their current interface provides the means to conduct traditional information retrieval; i.e., text-based search. Other re-

cent initiatives are, for example, described in.[11–13] One of the most famous CBIR systems available is IBM's QBIC[14] as launched in mid-90s. In January 1997, the IBM Corporate Community Relations project with the State Hermitage Museum in St. Petersburg started.[2] The goal of the project was to do much more than just provide technology to the Hermitage Museum. The project's aim was to transform how people around the world experience the Hermitage Museum and its collections.[2]

With respect to the database of the Rijksmuseum, there is a more pressing problem than the proper use of keywords. In general, modern information retrieval techniques provide excellent results[15] when two premises are satisfied: (i) a well annotated database is available and (ii) a good choice of keywords is made, which both fits the query in mind and the keywords present in the database. Using a limited database, such an approach can be highly successful. In contrast, in most situations, no well annotated databases are present in an unrestricted domain, which are queried by non-professionals, using non-optimal keywords.

The general public may not know the style of a painter, the period he lived in, or the painting's name. Often, a visitor does not even the exact name. How to approximate a name, using keywords? In such a scenario, the user cannot access the data to fill his information-need.

A professional brings his knowledge to bear about the artist (*e.g.*, name and country of residence), the object (*e.g.*, title, material(s), technique(s)), the dates, the acquisition method, and possibly will be able to use his associations. He can define detailed queries, which produce retrieval results with a high precision. However, how to find objects that evoke the same atmosphere or trigger the same emotions? How to find objects with a similar expression, although created using other techniques on different materials? Systems that can answer such questions should use intelligent information processing schemes, such as CBIR.

3 Presentation of results

In research on CBIR systems, most effort is directed toward the underlying search technique. The user interface of these systems receives little attention. In this respect, the current research differs from most other CBIR research. In general, in CBIR, the query is an image. The selection of this image can be done in two manners.

A user has an image that he wants to use as query that may not yet be available to the system, the interface needs a simple (upload) button to provide the system with the query image. Selection of the query image through browsing a large database, on the other hand, is complex. An image database often contains thousands of images; hence, manual browsing through the complete database is not feasible. So, the user interface has to both efficient and user friendly. Moreover, not only for the query definition but also for the presentation of the CBIR retrieval results a user interface is needed. This is important as people want to browse through the results.

We have chosen to base our user interface on scientific research. Until recently, there has been a lack in scientific research concerning the presentation of the results of a CBIR query. Van Montfort, De Greef, and Van Eijk[16] are the exception with their research toward 'Visually searching a large image database'. They found that manual browsing was most effective. A grid of 9-16 images per screen was found optimal for visualizing an image collection. Another interesting finding was that the difference in size between the (real-size) image (*e.g.*, a painting) that was seen and its thumbnail version, as it is available in the image database, does not hinder recognition. In order to facilitate efficient and user friendly browsing, the user interfaces are designed conform the guidelines of these authors.[16]

4 Online Content-Based Art Retrieval (C-BAR) system

The online Content-Based Art Retrieval (C-BAR) system presented here consists of three components:

1. User interfaces: query definition interface, the browsing user interface, and the user interface that provides the results.
2. The matching engine
3. The XML and image databases, connected through the image names.

4.1 User interfaces

The user interface consists of two sheets that can be accessed by selecting a tab. Each of the three sheets provide their own functionality. The online C-BAR system is available at: http://eidetic.ai.ru.nl/C-BAR/. Figure 3 shows four screendumps of the user interface of C-BAR.

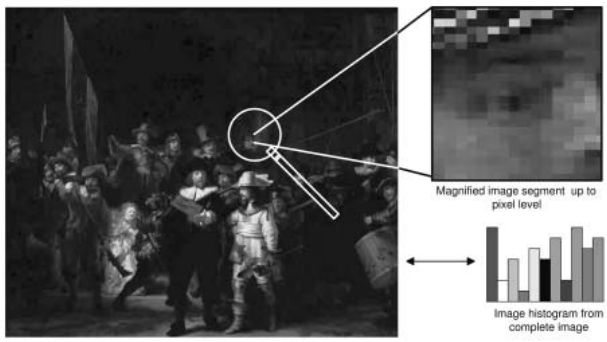


Figure 2. Feature extraction explained: Color analysis on pixel level, see the magnified image fragment. For each pixel its color is quantized and assigned to a color category. Next, the global color distribution of the image can be described as a vector, visualized as a histogram bin.

The first sheet provides the means to browse through the Rijksmuseum database and select an image as query, as is shown in Figure 3c. Also, a matching scheme (see Section 4.2 and 4.3) can be chosen and the number of retrieval results can be specified, as is shown in Figure 3d. After the query image and the matching scheme have been chosen and the number of images to retrieve is specified, C-BAR is ready to be used. The results are presented as follows: each retrieval result is an entry in the results user interface. On the left side of the row, a thumbnail version of the image is shown. Next to it, a verbal description of the image is provided, as available in the XML-database, as is shown in Figure 3d.

The second sheet provides the access to a traditional text-based search engine that can be considered as an improved version of the one currently available at the Rijksmuseum web site.1 Text-based retrieval can be used for specific information needs; *e.g.*, to look up which paintings of 'Rembrandt' are part of the collection of the Rijksmuseum. The results are presented in a grid of 15 thumbnails, as is shown in Figure 3a. When the user hovers over the thumbnails, meta information is shown.

4.2. Feature extraction

With features we denote characteristics of objects. In the context of images, features can denote colors, texture descriptors, and shape descriptors. In the current research, we use features derived from color and texture, where texture is the structure present in the image.

The global color distribution of an image is described using a quantized color space that divides the color space in color categories. Each pixel of an image is assigned to one of the color categories and hence, the amount of colors available in the image is reduced. For each color category can be determined how many pixels are assigned to it. This results in a feature vector of the color categories that can be visualized as a color histogram, as is shown in Figure 2.

For the quantization of color, two color spaces were used: the standard, but for humans perceptually not intuitive, RGB (i.e., Red, Green, and Blue) color space and the for humans perceptually intuitive HSI (i.e., Hue, Saturation, and Intensity) color space. The RGB color space was divided conform a 4× 4× 4and a 8× 8× 8 quantization scheme; i.e., each of the axes were divided in respectively 4 and 8 parts. With the HSI color space the Hue-axis (denoting the color) was divided with a higher precision than the other two axes with all three quantization schemes applied: 9× 6× 6, 12× 3× 3, and 18× 3× 3.

In addition to global color characteristics of image material, the local structure of pixels (denoted by their color) can be utilized to describe the image as a whole. Van den Broek and Van Rikxoort17 describe a few of the more intuitive methods for texture analysis. In the current research, we use the so called color correlogram for color induced texture analysis. Combined with the color histogram the color correlogram defines the 'Parallel-Sequential Texture Analysis' method, as is recently introduced.[17]

4.3 The matching engine

The matching engine has an interface that connects to the query definition user interface and to the user interface that presents the retrieved (matching) results. Through the query definition user interface, the matching engine receives the query image as well as the necessary parameters: the feature vector to be used and the number of results to be presented. Subsequently, the matching engine extracts the features (the color histogram and optionally texture features) from the query image and matches it to the image features extracted from the images in the database. The results that match are ranked based on their distance to the query image. Next, the images that match best are sent to the user interface that presents the results.

The distance between feature vectors and, conse-

quently, between the images, is determined by the intersection measure or the Euclidean measure. The intersection measure sums the absolute difference between two elements at the same position in different vectors. The Euclidean measure sums the squared difference between two elements at the same position in different vectors. For example, vector < 1 9 3 5 2 6 4 8 8 4 1 3 5 > and <7 2 8 3 3 6 5 1 9 4 7 2 7>, their intersection difference can be expressed as vector <6 7 5 2 1 0 1 7 1 0 6 1 2> and their Euclidean difference can be expressed as vector < 36 49 25 4 1 0 1 49 1 0 36 1 4 >. Consequently, for both distance measures, the distance is the sum of the elements of the difference vector, which defines the intersection distance as 39 and the Euclidean distance as $\sqrt{207}$.

4.4 The database

The database as provided by the Rijksmuseum will contain of the images and their annotations, specified in XML, as is shown in the example on the next page. The images are connected with their annotations through the image names, which are unique.

To speed up the online matching process, a database will contain pre-stored (or cached) retrieval results. This is a database in which all images are already matched with each other, using all feature vectors. Hence, the true retrieval process is done offline. In the online matching process, only the table with the matching results has to be read.

4.5 The system

The C-BAR system incorporates three functionalities: (i) text-based search, (ii) content-based search, and (iii) inspection of the results. Both text and content-based queries result in a list of retrieved images, where each entry in the list provides both the image and its annotation. The amount of retrieved images is chosen on forehand. The default setting is that 10 images are retrieved.

After a text or content-based query, the results can be inspected, using content-based techniques. In a matrix, the distances between all possible pairs of retrieved images are provided, as is shown in Figure 3b. Moreover, when the mouse pointer is placed upon a distance, the two images to which the distance refers to are shown. Hence, the relation between all images retrieved can be inspected. Consequently, the users can infer the system's working. Subsequently, users can learn to understand and even respect the system's performance.

```
<xml>
<record>
<title>
Maria de Witte Françoisdr (geb 1616).
Echtgenote van Johan van Beaumont. </
title> <maker>Mijtens, Johannes</maker>
<object.name>schilderij</object.name>
<material>doek</material> <description>
Portret van Maria de Witte Franc
¸oisdr, echtgenote van Johan van
Beaumont.
Staand, ten halven lijve, voor bomen
en een landschap.
Een bloem in de rechterhand. Pendant
van SK-A-746.
</description>
<image.reference>
http://www.rijksmuseum.nl/images/as-
sets/web/P-SK-A-747-00
</image.reference>
<priref>4548</priref>
</record>
</xml>
```

5 Discussion

Although this project has been as successful thus far, it is a pilot project. A proof of concept is provided that resulted in an online data mining system, which enables the access to the Rijksmuseum database, using either image content instead or verbal descriptions. Consequently, laypersons in the field of art, who are interested in the collection of the Rijksmuseum, can access its database in a simple and intuitive manner.

In the continuation of the current research, all aspects of the system will be improved. The user interface will be enhanced both with respect to its visual appearance and with respect to its functionality. So far, the image descriptions generated denote the global color distribution of the image material. In addition, texture and shape features can be extracted from image material. The latter two types of image information are also rich sources of information and should, therefore, taken into account. For this purpose, we will include the recently introduced schemes for color induced texture: 'parallelsequential texture analysis'[17] and the shape extraction and matching scheme as proposed in 'human-centered object-based image retrieval'[18], which both use the, on human perception based, eleven color categories color quantization scheme as proposed by Van den Broek, Schouten, and Kisters.[19]

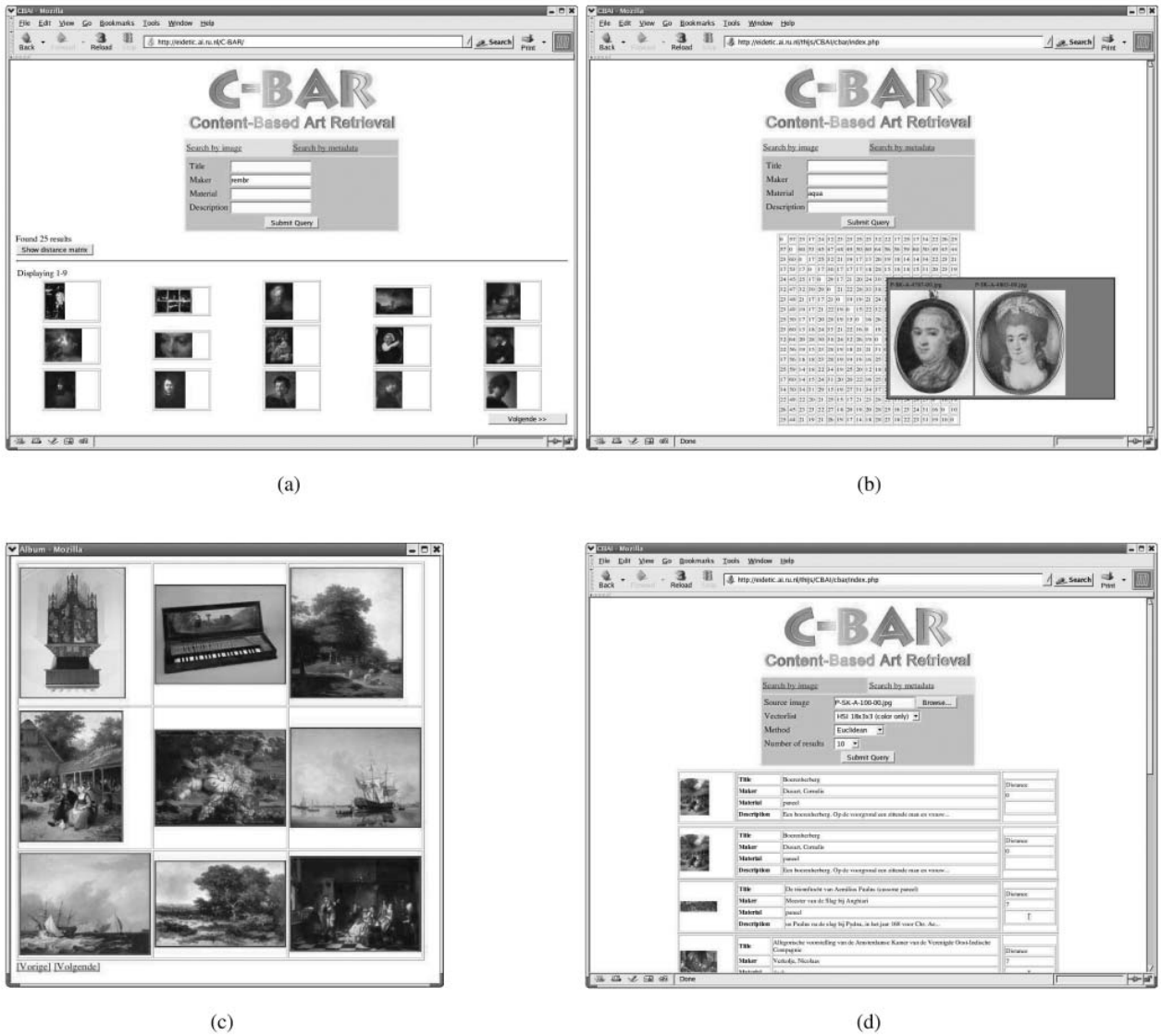


Figure 3. The online Content-Based Art Retrieval (C-BAR) system, available at http://eidetic.ai.ru.nl/C-BAR/. (a) The results of a text-based query, using meta-data. (b) The inspection of the results. (c) Browsing for an example image. (d) The results of a content-based query, using the HSI 18 × 3 × 3 quantization and the Euclidean distance measure.

This color quantization scheme mimics human color quantization and, hence, satisfies the aim to provide intuitive results to the system's users.

In general, for complex (multi-media) data-mining systems, user profiling is adopted as a paradigm. In contrast, we propose to include artist profiling instead. Subsequently, features as can be found in artists' work are identified (*e.g.*, number of colors, contrasts, texture) and utilized in such a profile. The resulting prototype feature vector could enhance the C-BAR system substantially.

In time, ontology-based knowledge representations should be used to define structure in the database instead of merely flat XML data representation. Moreover, cooperative annotation as envisioned and applied by Schomaker, Vuurpijl, and De Leau[8] should be applied. Of special interest would be to compare how laypersons and experts annotate the material. Learning algorithms (*e.g.*, neural networks) or techniques as singular value decomposition can extract underlying (implicit) dimensions of judgment[15], which are common to both groups.

C-BAR is still in development; however, its current version showed to work successful. It provides access both to an advanced text-based search engine for experts and to (ten configurations for) content-based retrieval that do not require any domain knowledge. Moreover, the means to understand its working are provided. Hence, a unique system is introduced to access, enhance, and retrieve knowledge from digitized art collections.

Acknowledgements

We thank Xenia Henny and Kees Schoemaker for their cooperation. They provided us the database of the Rijksmuseum. The Dutch organization for scientific research (NWO) is gratefully acknowledged for funding the ToKeN projects Eidetic (nr. 634.000.001) and VindIT (nr. 634.000.018), in which this research was partly conducted.

References

1. Rijksmuseum, 'Research,' URL: http://www.rijksmuseum.nl/wetenschap/ [Last accessed on April 21, 2005].
2. IBM, 'The state hermitage museum,' URL: http://www.hermitagemuseum.org/ [Last accessed on May 29, 2005].
3. L. Vuurpijl, L. Schomaker, and E. L. van den Broek, 'Vind(x): Using the user through cooperative annotation,' in Proceedings of the Eighth IEEE International Workshop on Frontiers in Handwriting Recognition, S. N. Srihari and M. Cheriet, eds., pp. 221–226, IEEE Computer Society, Los Alamitos, CA, (Ontario, Canada), 2002.
4. T. Kato, 'Database architecture for content-based image retrieval,' in *Proceedings of SPIE Image Storage and Retrieval Systems*, A. A. Jambardino and W. R. Niblack, eds., 1662, pp. 112–123, (San Jose, CA, USA), February 1992.
5. V. N. Gudivada and V. V. Raghavan, 'Content-based image retrieval systems,' *IEEE Computer* 28(9), pp. 18–22, 1995.
6. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, 'Content-based image retrieval at the end of the early years,' *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12), pp. 1349–1380, 2000.
7. Y. Rui, T. S. Huang, and S.-F. Chang, 'Image retrieval: Past, present, and future,' *Journal of Visual Communication and Image Representation* 10, pp. 1–23, 1999.
8. L. Schomaker, L. Vuurpijl, and E. de Leau, 'New use for the pen: outline-based image queries,' in *Proceedings of the 5th IEEE International Conference on Document Analysis*, pp. 293–296, (Piscataway (NJ), USA), 1999.
9. C. Jörgensen, 'Access to pictorial material: A review of current research and future prospects,' *Computers and the Humanities* 33(4), pp. 293–318., 1999.
10. Instituut voor Nederlandse Geschiedenis, 'Web site Instituut voor Nederlandse Geschiedenis (ING),' URL: http: //www.inghist.nl [Last accessed on May 30, 2005].
11. F. Hernandez, C. Wert, I. Recio, B. Aguilera, W. Koch, M. Bogensperger, P. Linde, G. Gunter, B. Mulrenin, X. Agenjo, R. Yeats, L. Bordoni, and F. Poggi, 'XML for libraries, archives, and museums: The projects covax,' *Applied Artificial Intelligence* 17(8), pp. 797–816, 2003.
12. N. Davenport, 'Council on library and information resources.' URL: http://www.clir.org/, [Last accessed on July 14, 2004].
13. J. Trant, *Image Retrieval Benchmark Database Service: A Needs Assessment and Preliminary Development Plan*, Archives & Museum Informatics, Canada, 2004.
14. M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D.

Petkovic, D. Steele, and P. Yanker, 'Query by Image and Video Content: The QBIC system,' *IEEE Computer* 28(9), pp. 23–32, 1995.
15. E. C. M. Hoenkamp, 'Unitary operators on the document space,' *Journal of the American Society for Information Science and Technology* 54(4), pp. 319–325, 2003.
16. X. A. N. D. R. A. van Montfort, P. H. de Greef, and R. L. J. van Eijk, 'Visually searching a large image database: Manual browsing versus rapid visual presentation,' [In preparation] , 2005.
17. E. L. van den Broek and E. M. van Rikxoort, 'Parallel-sequential texture analysis,' in *Proceedings of the 3rd International Conference on Advances in Pattern Recognition* (ICAPR2005), S. Singh, P. Perner, and C. Apte, eds., [accepted], 2005.
18. E. L. van den Broek, E. M. van Rikxoort, and Th. E. Schouten, 'Human-centered object-based image retrieval,' in *Proceedings of the 3rd International Conference on Advances in Pattern Recognition* (ICAPR2005), S. Singh, P. Perner, and C. Apte, eds., [accepted], 2005.
19. E. L. van den Broek, Th. E. Schouten, and P. M. F. Kisters, 'Efficient color space segmentation based on human perception,' [submitted].

Backing the Right Horse: Benchmarking XML editors for text-encoding

*Thijs van den Broek, Frans Wiering & Roelof van Zwol**

This paper is a condensed version of Thijs van den Broek's master's thesis at the University of Utrecht. For a complete account of this research we refer you to this thesis [Broek, 2004]. The research was conducted in spring 2004 at the Oxford Text Archive (OTA) based at the Oxford University.
The OTA is one of the largest digital text archives in the world. Its main goal is to store texts that are digitized by scholars so that other scholars can reuse these texts. Therefore they advise scholars to use XML for text-encoding but little is known about which editors are best suitable for encoding the texts. This research was carried out to find the best suitable XML editor for text-encoding.
By creating a benchmark, a generic and re-usable solution was found to pick the best suitable editor for different types of users irrespective of the document model (DTD, Schema) being used. This paper describes the creation and use of the benchmark that consists of three parts (features, usability and support). The benchmark can be used to select the best suitable editor for four types of users by ranking the editors from most suitable to least suitable based on the requirements of each user group. After the ranking of the editors, user tests are held to find the best suitable editor for their group.

Problem statement
The OTA wants to advise XML users about which editor they should use. For their advice to scholars they need a list of requirements from scholars and a selection of editors that comply with these requirements. Therefore the problem from the OTA is the following.

What are the requirements for an XML editor to encode texts given an XML encoding schema and a type of user? Which editors comply with these requirements?
The solution to address this problem is to create an instrument that can evaluate XML editors based on the requirements. This will result in a score for each of the editors on which they can be ranked. This results a second problem.

Can we create a benchmark that addresses the first problem?

In the following paragraphs we will argue that we have created such a benchmark.

Method
Numerous benchmarks are available to evaluate different types of products, including software products. However, benchmarks for XML editors were not found. Therefore the scope of the literature was ex-

*Centre for Content and Knowledge Engineering, Department of Information and Computing Sciences,Utrecht University

tended to 'evaluation of software'. This query resulted in many checklists for various software products, but also some research on evaluating software.

One paper on evaluation of software was particularly interesting because it criticized the commonly used approach in software evaluation using numerically weighted lists of criteria. Instead of a numerically weighted list the paper describes a qualitative weighted list of criteria [Baumgartner, 1997]. Instead of taking values from 1-5 or 1-10, symbols were used to describe whether a criterion was: E (=essential), * (=very valuable), # (=valuable), + (=marginally valuable) or 0 (=zero). The symbol that was given to a criterion determined the maximum score a product could receive on that criterion. So if a certain criterion was weighted # (valuable), the symbols given to a software product on that criterion in the evaluation could only be: #; +; or 0. This type of software evaluation has a major advantage over usual checklists of criteria. When software is evaluated based on equally weighted criteria, some criteria are overvalued in the benchmark. By evaluating software based on criteria that have different weights this overvaluing of criteria is prevented.

The International Organization for Standardization (ISO) has also published papers on the evaluation of software. Its standard ISO 9126 (2001) describes a framework of characteristics that should be tested with a benchmark. One characteristic of the quality model in ISO standard 9126 is Usability, an important aspect of software. A good instrument for evaluating usability is that of the heuristics of Nielsen [Nielsen, 1993]. The advantage of this evaluation method is that it is split up in general criteria like Consistency, Feedback, Help and Documentation, etc. With these heuristics it is possible to judge if an XML editor satisfies the usability guidelines.

The ISO quality model, the heuristics of Nielsen, and the qualitative approach on evaluating software from Baumgartner method are combined for the creation of the benchmark for XML editors for text-encoding. The criteria in the benchmark are then weighted by users.

Second stage in the research was interviewing users of XML editors to find out what they need for their XML editing work. Therefore 14 people with different perspectives on the subject were interviewed. The two main questions concerned important features of XML editors and different user groups. The answers were used to create a list of features and a distinction of user groups.

Benchmark

After the literature search and interviews with users, enough information was gathered to start creating the benchmark. It consists of three parts.

– Features
– Usability
– Support

The first part consists of criteria for features of the editor. Users from the different user groups are asked to assign weights to these criteria. With this approach, users can determine which features are most important based on their work and skills. The second part consists of criteria for the usability of an editor. This is a separate part of the benchmark because of the importance of usable software programs. Users do not evaluate these criteria because they are important regardless of the type of text-encoding. The maximum value an editor can receive on the usability criteria was 'Very valuable'. The last part consists of criteria for the ease of support for the editor. This category is developed to check whether the editor is easy to support by the support departments. Users do not evaluate these criteria because they are important for support departments and not the users themselves. The maximum value an editor can receive on the support criteria was 'Very valuable'. The complete benchmark can be downloaded from the I-Cult web site [2004].

User groups

The interviews were also used to define different groups of users in XML editing. People were asked to create distinctive groups. Based on the interviews the following list of user groups was created.

Group 1: Advanced users in XML technology

Advanced users in XML technology publish and research complex documents. Therefore, they encode the text very thoroughly working in the XML itself. They also create and edit schemas and style sheets. This requires a lot of knowledge about different XML technologies.

Group 2: Advanced users in XML encoding

Advanced users in XML encoding are similar to Group 1 but do not create and edit schemas and style sheets. This group has a lot of knowledge about XML encoding.

Group 3: Intermediate users in XML encoding
Intermediate users in XML encoding are familiar with XML but find it easier to use a WYSIWYG editor. This group prefers to work in a view that shows only tags or no tags at all rather then in the XML code itself. The documents they inherit may be invalid.

Group 4: WYSIWYG authors
WYSIWYG authors are not interested in XML, but are required to work with it by their organisation. They want to work in an environment similar to word processing. They do not want to see the XML itself.

Weights for the criteria

When the list of criteria was complete XML users were asked to complete a questionnaire asking them in which group they belonged and how they evaluated the list of thirty features. They were asked to assign one of the following values to the criterion:
– Not valuable
– Marginally valuable
– Valuable
– Very valuable
– Essential

Evaluating the editors

When the weights were set, the next step was to evaluate the features of the editors. This was done by the first author of this paper. Each editor was checked for the implementation of the features listed in the criteria. The minimum score an editor could get for a single criterion was 'Not valuable' and the maximum was the average value that the group of users had decided for that criterion. When every editor was evaluated a ranking was made, sorting the editors from best suitable to least suitable editor for each of the user groups. After the evaluation of the features the usability of each editor was evaluated. Each editor was checked against the criteria for usability. This resulted in a new ranking of editors based on usability. Final evaluation step was the evaluation of the ease of support for the editor. After checking the editors against the support criteria, a ranking was made based on the ease of support of the editors.

User testing

When the evaluation of the editors was completed a shortlist of editors was created based on the rankings for features, usability, and support. For each of the user groups a group of editors was selected for user testing. The editors on the shortlist for each group were evaluated by a sample of three users from that group. The evaluation consisted of performing a few tasks per group in the selection of editors. By giving marks for the performance of the editors on the tasks and ranking them, the users selected the best suitable editor for their group.

Results

Weights for the criteria
The questionnaire for setting weights for the criteria was send to several XML user lists and colleagues within the University. A total of 92 respondents completed the questionnaire of which most were in Group 1 (n=56), the other groups contained less respondents: Group 2 (n=12); Group 3 (n=9); and Group 4 (n=16).

The results of the completed questionnaires were somewhat disappointing. The given weights to the criteria did not create a big distinction between the groups. In each of the user groups almost every criterion was evaluated 'Valuable' or 'Very valuable'. This was probably because the users did not have to choose between features but could pick all of them. Fortunately, some features did show the distinction between user groups.

Evaluation of the editors

When the weights of criteria were set, the next step was the evaluation of the editors based on the criteria in the benchmark. This was done with the same values as used in the weighting of the criteria with one exception. The value 'Essential' could not be given as a value for implementation of a feature in a product and was therefore left out of the range of values given to a product. Therefore the values that were given to the products on criteria in the sections of features, usability, and support the editors were the following.
– Not valuable
– Marginally valuable
– Valuable
– Very valuable

Ranking of editors per group

After the evaluation of editors, the next step was to create a selection for each user group based on the weights of the criteria and the evaluation of the edi-

tors. This has to be done by an expert. In this research, the first author of this paper made the selection. If a user group (for instance Group 1) gave the criterion 'Preview the formatted text' the value 'Marginally valuable', this meant that this was the maximum value an editor could receive for this feature within this user group. If another user group (for instance Group 4) gave the value 'Very valuable' to that same criterion, this meant that that was the maximum value an editor could get for this feature. So if for instance the XML editor XMetal scored 'Valuable' in the evaluation of the editor it received the value 'Marginally valuable' in Group 1 (which is the maximum value) and it received 'Valuable' in Group 4 (which is the actual value of the editor).

By using this approach used by Baumgartner [1997] for the evaluation of software, the selection of editors for the groups should have differed based on their evaluation of the criteria. However, as was mentioned earlier, the problem was that there was only a slight difference in the evaluation of criteria per group. Therefore, the ranking of the editors was also just slightly different per group. Since most features were evaluated as 'Valuable' or 'Very valuable' by each of the groups, the editors that had the most features implemented topped the ranking in each of the groups.

Besides the ranking of editors based on features, two other rankings were made based on usability and support possibilities. These two rankings were used as filters on the ranking that was already made based on features. In the usability ranking the word processors Microsoft Word and OpenOffice topped the ranking because of their familiarity with users and easy user interface. In the support ranking the editors that topped the ranking were editors that were highly configurable or portable, for instance Epic Editor and Emacs.

After the ranking of the editors, the next step in the evaluation was to select editors considering the three different rankings. As mentioned above the problem of the lack of distinction between the groups had to be resolved. If the groups had only evaluated the features that were actually valuable to them with 'Valuable' this would have made a distinction in the weights of criteria between the groups. Then there would have been a clear distinction between the rankings of editors per user group and the top ranked editors could have been selected for user testing. Because the weights of the criteria by users did not make that distinction, this distinction had to be made by the first author of this paper.

To create the distinction, more weight was given to the features that actually did create a distinction between the groups. The important features for a particular group got more weight in the selection. For instance, Group 1 needs to have functionality for writing style sheets and schemas, therefore only editors that have this functionality were chosen for the shortlist. Group 3 needs to have a code view mode as well as a WYSIWYG mode, so editors that have both modes were chosen. Group 4 needs an editor that has close resemblance with a word processor, so only editors with a WYSIWYG mode were selected.

Based on this information the following editors were selected for user tests by users from the different groups.

Table 1: Selected editors for user testing per group

Group 1	Group 2	Group 3	Group 4
Oxygen	Oxygen	Epic Editor	Authentic
Emacs	Exchanger	Oxygen	Serna 2.0
Exchanger	Emacs	Serna 2.0	XMetal
		XMetal	OpenOffice

User tests

Because of our intervention it was possible to start testing the selected editors by users. Users from each of the groups were asked to perform some quick and simple tasks with the selection of editors. Twelve users (three per group) completed the user tasks. They were asked to give marks to the editors from 1 to 5 (where a 5 was the maximum score) for suitability and user-friendliness in performing the tasks. They were also asked to give an overall score for the editor with a mark from 1 to 5.

At the end of the user tests, users were also asked to rank the editors from best to worst editor. They choose the following editors as best suitable.

– Group 1: Oxygen
– Group 2: Exchanger XML
– Group 3: XMetal
– Group 4: XMetal

Conclusions and future research

Considering the first problem we researched, we can conclude that we have found the requirements for an XML editor to encode texts given an XML encoding schema and a type of user. We also found the editors that comply with these requirements. These re-

quirements and user groups were gathered during interviews. With a questionnaire the weights for each requirement was set by the user groups themselves. Based on the weights and the evaluation of the editors by an expert the editors were ranked for each group. These editors were then tested by users from that group. This resulted in a list of editors that comply with the requirements. However, the problem with this approach was that most users evaluated the criteria as 'Valuable', even when they were not going to use it. This caused similar weights for the criteria in each group. Therefore, the same editors topped the ranking in each of the groups. These were not the editors that were most suitable for that group, but the editors that had the best overall score based on the features. Forcing users to choose between the different criteria instead of evaluating each criterion separately will solve this problem. That will result in a better distinction between the user groups. For instance, advanced users will then likely choose a built-in XSLT processor over working in WYSIWYG mode whereas WYSIWYG users will likely choose a WYSIWYG mode over the XSLT processor.

Considering the second problem we researched, we can conclude that we have created a benchmark that addresses the first problem. The benchmark makes it possible to evaluate completely different XML editors and rank them based on the requirements of different user groups. However, because of the problems described earlier the benchmark is not perfect yet. Despite the problems, the benchmark is a good instrument for choosing the best suitable XML editor per user group. Where other benchmarks of software are often just checklists with features that are checked or unchecked this benchmark gives the user insight in what requirements are important for his text-encoding work and which editors comply with those requirements.

Future research

As mentioned earlier, the main problem with the benchmark is the fact that users are not forced to make a choice between the criteria and therefore evaluate all the features as valuable. When a second version of the benchmark is created, this problem has to be addressed. The first step in creating a new version is the creation of a new questionnaire. Users from the four groups have to be asked to evaluate the thirty features again with the remark that they can only use each of the values six times, for instance (thirty features divided by the five different values). This ensures that the users choose which criteria are most important. Other sets of limited values are also possible as long as it forces the user to make choices between the features.

Re-use of the benchmark

The benchmark was created so that it can be re-used by others for future evaluation of XML editors. It can be used regardless of the document model being used by the user. When new editors are developed they can also be evaluated with this benchmark. This evaluation has to be conducted by someone with knowledge of XML, XML editors, and the criteria listed in the benchmark. The results of the evaluation of editors can then be added to the evaluation results of other editors. The next step is the adjustment of the evaluation of the editors based on the evaluation of the criteria per user group. This will result in four different rankings of editors, one for each group. Users can test the most highly ranked editors by performing a few tasks with them. This will result in a ranking of best suitable editors for text-encoding for their group. The complete benchmark can be found at the I-Cult web site [2004].

References

Baumgartner, P. (1997). Methods and Practice of Software Evaluation. The Case of the European Academic Software Award. *Proceedings of ED-MEDIA 97 – World Conference on Educational Multimedia and Hypermedia,* Charlottesville.

IEEE STD 610 (1990). *IEEE Standard Glossary of Software Engineering Terminology*.

ISO/IEC 9126-1 (2001). *Software Engineering – Product Quality – Part 1: Quality Model.*

ISO/IEC 9126-2 (2002). *Software Engineering – Product Quality – Part 2: External Metrics.*

ISO/IEC 14598-1 (1999). *Information Technology – Software product evaluation – Part 1: General Overview.*

Nielsen, J. (1993). *Usability Engineering*. San Diego, Academic Press.

Broek, T. van den (2004). *Backing the Right Horse: benchmarking XML editors for text-encoding. Utrecht, Utrecht University.* Retrieved May 25, 2005, from http://www.cs.uu.nl/research/projects/i-cult/Publications/ BenchmarkXMLeditors_Broek.pdf

Digital archives in a virtual world

Regina Brunnhofer & Ingo H. Kropač

In the last years the role of the FCR has changed slowly from an editing project to a small scientific organization, which aims at a hybrid status between archiving and editing as well as software development and longtime preservation of digital data. The FCR sees its position at the intersection between history, information science, documentation and archival studies. Just aspects like reliability, preservation and availability have become more important than simply finding the correct text. Computer based methods and techniques of textual representation and (analytical) processing have overcome the space of linearity, new concepts change traditional scientific behaviour. Thus we will abandon the software-driven approach in favour of a data-driven concept and we propose an unusual way in documenting, a 'work in progress' which *per definitionem* can never be completed. Following the out-lined methods and concepts we try to make our contribution to the 'new' virtuality.

What historians really do since centuries when they write or explain history, is creating their own viewpoint of the past, is producing virtual worlds. Sometimes these worlds are very close to facts, sometimes they represent only the idea of experts and non-experts in history. Looking at some 'historical' games and simulations, the border between virtuality and reputable research becomes more and more blurred. The development of the last decades is towards the so-called 'fun-society' favoured virtual worlds, where it is difficult to distinguish between facts and fakes.

Also the traditional roles and tasks of archives have changed due to the change to the information society. Many records are only stored in electronic form, a lot of systems support paperless transactions and the issuing in a paper form is only one possible representation of the original document. At the same time, originally analogue data is getting digitised. Future generations of historians will face the same problems like their predecessors after the 30-Years-War: How authentic are the sources and documents, they want so use for their explanations? But in contrast to the situation about 350 years ago, our successors will have major difficulties to find correct answers.

1 General considerations

What is an archive? Generally speaking, archives are arranged collections of documents. As a rule, documents which are meant to be archived are transferred from their 'producers' to the archives, where the emphasis is on the collection and long-term storage of archival records.

In this context, the term 'electronic information' is mentioned very often, an item which has various meanings: Firstly, electronic information can mean data, originally mostly in paper form, which either by data-entry or scan or optical techniques has been transformed into electronic data, for example, scanned photographs and documents. Secondly, electronic information is, of course, also material which is 'born digital' and has no analogue equivalent, such as photographs taken by a digital camera. In other words,

*Historical Information and Documentation Science, Graz University

electronic records, files and documents are not physical, but virtual units.

1.1 Theses on terminology and typology

But what is a digital archive? First of all, it has to be put straight that a 'digital archive' can have several flexible characteristics: Firstly, digital archives may be a digital representation of existing archives, together with appropriate detailed finding aids for individual record groups or collections, repertories and digitised data. Secondly, a digital archive can form a something completely new, a virtual archive. Thirdly, data archives also belong to digital archives. Data archives are collections of digital information for the purpose of secondary analysis. Unlike the first type of digital archive, the information within this data archive has no analogue equivalent.

What all three types have in common is that digital archives should see their main task in providing easy access to well-documented digital resources. Access is assumed to mean continued, ongoing usability of the digital resources. In cyberspace, it doesn't matter where data is preserved and by whom, but what counts is to know where the data is (Doorn, Research 108). Archiving means more than the durable storage of digital information on a data medium. It also includes the preservation of the durable availability of digital resources. Long-term preservation refers to the selection, ordering, cataloguing, development and maintenance of digital data and/or data media. This also includes the development of strategies to ensure that the data of digital resources is conserved. Digital archives should see their role as mediators between the producers of digital information and the users.

Prerequisites for digital archives

A requirements analysis is crucial for the production of a digital archive. What should be kept for future generations? Will the information we bequeath satisfy their requirements? So archiving is a balancing act because, on the one hand, you have to avoid storing too much information and, on the other hand, nothing important should be lost.

This problem is linked very closely to user group analysis. Virtual archives must be exactly co-ordinated with the needs of the users: do they have previous experience, what are their aims and expectations? Therefore, the quality of the user interface is one of the most important things: the archives' on-line finding aids should include an administrative history or biography, a scope and content note, acquisition information, processing information, series descriptions and folder lists. In other words, users should be able to easily orient themselves in digital archives and receive the information they are looking for, even though they may need some guidance. If these factors are not considered, you run the risk of building a 'virtual cemetery'.

In terms of data security, you have to bear in mind that there must be appropriate tools to prevent unauthorized access. Moreover, digital archives have to provide authorized access to the data.

The 'success' of a digital archive doesn't depend the least on the handling of metadata. It is of essential importance to have a sufficient description of the data and the metadata, which describes the functioning of the institute that produced and processed the data. This additional digital information is primarily intended to make the collection more accessible. But it is also an important tool in managing the archives.

1.2 Advantages, constraints and strategies

Incorporating these preliminary ideas, digital archives offer incontestable conveniences: First of all, the users' access to sources is not bound to a location. They act in a virtual reading room where a huge amount of data can be viewed. Furthermore, the dynamic of digital archives provide an opportunity for rapid modifications and supplements. Thus, a variety of different questions can be solved rapidly. To finish the boring discussion on the 'best order' of archives, digital archives enable the simultaneous displaying of different ordering systems, *e.g.* the provenance- and subject-oriented principles.

But there has to be careful consideration of what should be preserved and filed. Records Management, in terms of the scraping of documents, is a crucial point with regard to digital archives. The British National Archives, for example, has already accepted this necessity and compiled guidelines to this effect.

Naturally, there are also constraints and disadvantages which have to be considered: the ephemerality of data media is one of the most serious. If you consult the internet as an example, then this problem shows up very clearly. The web is the biggest document which has ever been created, but the average life-span of a site amounts to 44 days. This loss of information is tremendous, especially because some of this material is unique and is thus irrecoverably lost (Roetzer,

Vergessen [passim]). Not in the least because of this, strategies for reliable long-term preservation need to be designed. Migration, for instance, involves the transfer of electronic records that can only be correctly read and interpreted to a new technology platform by legacy computer hardware and software. This transfer requires the design of gateways from the legacy system to the new technology platform and the writing of special purpose codes or programs to transfer the records and the software functionality. Typically, the migration of electronic records involves a number of complex issues, is very costly, and requires more time to complete than is projected.

Emulation, a further possibility in this context, means that data is kept in its original format and the necessary hard- and software is accordingly adapted so that the old formats can still be read. To express this differently, it is developing techniques for the imitation of obsolete systems on future generations of computers.

Some may see the fact that the job description of archivists will completely change as a drawback of the digital world. Core competencies such as acquisition or storage persist, but, in the digital age, this can stand for the development of a digital edition. Hence, the archivist increasingly becomes an information manager and 'processor' of data. Working in digital archives also involves a lot of teamwork, which is not bound to a specific location, but can take place in a decentralized way.

Digital archives have varying archiving systems, description systems and ways of accessing and using the data. In other words, long-term preservation is not standardized, but recommendations for the structuring of digital archives do exist:

– In 2001, the European Union Commission started the '*Lund-Initiative*', a catalogue of guidelines which aims to align digitization projects at the European level. Thereby the emphasis is on locating digital resources and content and its presentation.
– The RLG-Group, which is an international federation of more than 150 libraries, archives and museums, offers information exchange, discussions and suggestions for solutions concerning those organisations which have to meet the challenge of the digital era. Its handbook, the *'Trusted Digital Repository Report',* has worked out the most significant criteria of digital archives.
– One of the recent initiatives is the nestor – task force *'Vertrauenswürdige Archive – Zertifizierung'*, started in December 2004. The group aims to compile a checklist and develop an engaging certification procedure for digital archives. The first results are to be presented in December.

2 The Fontes Civitatis Ratisponensis – a case study?

2.1 The state of the art critically reviewed

The name *Fontes Civitatis Ratisponensis (FCR)* described on the one hand the effort to compose editions of the documental sources of the *Reichsstadt Regensburg* from the Middle Ages to the early modern times; on the other hand a concept of synergetic cooperation between archivists, historians and the exponents of the so called historical basic sciences (*Historische Hilfswissenschaften*) to documentate and retrieve our (written) cultural heritage. In order to reach as many interested parties as possible, local computers as well as networked systems are used as a carrier for the documentation of highly structured sources and web-based facsimile editions. In addition to those parts available online, CD-ROMs are produced to put the performance of the (static) WWW versions at the user's disposal on stand alone computers.

The FCR in their present size are the product of several research projects and the result of a co-operation between these projects and the *Amt für Archiv und Denkmalpflege* of the city of Ratispone. Besides the theoretical, methodical and (data) technical project contents the FCR contain also independent sub projects, that either deal with a closed source stock or with superordinated tasks such as software developments or the build up of the virtual archive, which is the most recent subproject of the FCR. The goal is to provide the written tradition of the city of Ratispone independent of its physical storage. It will be made accessible via the internet from all over the world. This is a big advantage because the sources of the history of Ratispone are scattered among several archives. Only a small part is accessible for the historical research in the form of printed editions. Moreover, the virtual archive might put an end to the discussion about the observance of the principles of archival arrangement (provenance or original order). This, because in the virtual archive many principles of arrangement can be used in parallel without changing or destroying the stocks in which these sources are stored. At present

approximately 1100 sources or stocks, respectively are registered and accessible in over 16.000 images.

The stocks of the particular sub projects show a different degree of processing. Some of them have been completed according to the methods of the FCR and can now be used for various kinds of publication media or have been already published. In some other parts of the subprojects there are some gaps in the inventory, or some processing steps could not be finished which would have been essential for the chosen concept. Because of the high number of sources which have been integrated in the whole system by now it will become the central matter to be able to use the total development in the context of the so called virtual archive. This is planned for the overall system – this is both for the users of the FCR and for the internal project system administration and processing. From the enormous expenditure of work invested in the processing of the particular stocks it appears to be understandable that there has been no space for a 'stock spanning' development in the preceding projects. That means that presently no 'stock spanning' catalogues or finding aids have been developed and so also an effective instrument for efficient controlling and evaluation is still missing. This lack not only concerns the management of the project co-operations and all colleagues but also the potential users of the accessible system modules. For the users it is important to know whether they view a 'complete' edition in the meaning of a traditional printout or whether they have accessed an early version of an edition that is still in progress.

Therefore, the transparency of the progress of the work as well as the transparency of the content-related processing represent a desideratum for the adequate presentation of the results of the project series to the special public. Knowledge of the processing of the particular stock must have influence on the total development in a cumulative way, in which the particular state of processing of the particular stock as well as the integration in the complete system have to be documented. That means that all parts of the core system that are not accessible at present will be available for the different control authorities and user groups by appropriate interfaces via the internet. The necessary software solutions for that will be realized in the present project; to use this software accordingly is a precondition for an evaluation system and the scientific quality of the results.

A last problem area is attached to the one just mentioned. The acceptance of digital resources such as produced by the *Fontes Civitatis Ratisponensis* often suffers not only from a missing quality management, but primarily from the fact that the World-Wide-Web is considered not to be very reliable. Therefore, the capabilities of digital resources for the humanities are not exhausted – particularly in the World-Wide-Web – and they still don't have the scientific acceptance they deserve. There are many reasons for this: There are no editorial boards or quality criterions as there are usually for the print media. In addition, presentations on the internet are not reliable because it is not guaranteed that they are available permanently and thus quotable. Therefore, the acknowledgment of WWW resources in scientific community is hardly existing.

For named reasons it is necessary to establish certificated servers which are approved by (humane and culture-scientific) specialized public. To plan and to realize such servers, preliminary conceptual work must be done in the areas of version control, long-term availability, data migration, quotability and 'archiving', technical conversions in prototypical form must be created. Then it will be possible to convince scientific institutions from universities to international associations to set standards on the basis of such preliminary work and to offer services in this area.

For many reasons the above mentioned cannot, in addition, be part of projects such as recent one; however, the experiences made shall be placed at the disposal of such a server project. The responsibility of a project series like the FCR for the long-term 'survival' of their results is imperative. Print-outs indeed can hold final results, but not the way how they got there. This means that in order to achieve the methodological objective – such as transparency and comprehensibility – all data at all levels of the processing are to be 'backed up' which guarantees a long-term availability and system independence.

2.2 The general aims – modularity, transparency, trans-disciplinarity

A principal purpose of the FCR is to combine reliable techniques in scholarly editing with recent information technologies and the scientific-theoretical approach. This means interdisciplinary work and transparency of a widely defined editing process by the application of formal methods combining rule-based automated procedures with the application of an editor's special knowledge.

Aim is the production and successive extension of an integrated information system. On the one hand

this system serves for archiving and textual development of original historic sources, on the other hand, however, it also provides extensive aids for analysis. As a result this integration system broadens the functionality of print-media based editions: In addition to text editions, as they are defined in the traditional understanding, it will be necessary to provide more and more structured information for users; the more 'serial' (or structured) the sources the more structured the information. Therefore it is not only necessary

– to illustrate the entities which need to be documented (historic sources and objects) in different representation forms (image, text and facts databases),
– to distinguish precisely between the documentation of entities and the knowledge about them accurately as well as to handle this expert's knowledge adequately (declarative and procedural knowledge representation),
– to connect all images and knowledge bases that have evolved by applying the above mentioned knowledge representation with classical data base applications,
– as well as to make the locally implemented overall system or only parts of it world-wide available for international computer networks using suitable software.

As a matter of facts it is also necessary to introduce new structural viewpoints in the definition of edition to be applicable for all possible types of sources, *e.g.* serial sources or miscellaneous office books with complex and very heterogeneous structures.

In such a system there need to be different ways of access depending on the user groups: For casual users, editions are exported from the basic system to a web-system and are presented via the Internet or on CD-ROM. The requirements of advanced users will be satisfied by assistant-driven CGI-based interfaces. Expert users are able to directly access released parts of the system objects themselves.

For the time being, the core system of the FCR is based on a software called 'kleio', a package which was created specially for the needs of the historic sciences. Also other big projects in reputable research institutions, archives and museums undertake considerable effort to provide material which can be summarized under the broad term of 'cultural heritage' in the WWW. In many cases the result of such efforts are static presentations of WWW pages. Indisputably, the usability of historic sources on the internet is quite enlarged if tools for a better accessibility of specific information are implemented. However, historic sources have the biggest use in the internet, when they are not only explored and put in the internet, but can be analysed according to different viewpoints. For the development as well as for the analysis of sources data base-supported applications and procedures are practically inevitable. For this purpose a modular organization of objects was developed which constitutes the overall system by the sum and the connections of all modules; various databases, knowledge representations and additional software tools belong to these objects.

2.3 Special aims and actual work

As we have lined out in the previous chapters two main goals for actual work can be identified:

1. The virtual archive which has been already realized in central areas during the running project needs to be further developed by the integration of a 'stock-spanning' whole development and an evaluation system.
2. The completed and standardized data of the overall system should be migrated just as all other declarative components of the knowledge representations in a largely system-independent format which follows industrial standards.

Persuing these main goals it should be ensured that the quantity and quality of the stocks which have been already incorporated in the system will be controlled and backed-up in the long term.

In particular, the virtual archive needs further substantial development. The goal of the archive is to provide a basis for sources of different stocks without changing or destroying these originating stocks. This means, that all existing representations of sources – the documents of the *Reichsstadt Regensburg*, the documents of the *Regensburger Almosenamt*, the *Bürgerbücher*, the city books, the *Cameralia* etc. – and their developments must be accumulated and united so that they are available via WWW for complex and 'stock-spanning' queries. In order to achieve this availability, the core systems of the respective sources have to be combined with each other. In addition they have to be further developed to a complex virtual archive by the help of catalogues of description systems and indexes (ontologies). Within this framework of the virtual archive it is necessary to develop metadata not only from a separate stock, but also from processing steps. These metadata will then be integrated into the system. This enables users to follow the processing

steps in detail such as: which person has worked with which stock or individual source; at what time this work has been done; when a release can be expected. In addition, there must be the possibility to integrate the competence of the scientific advisory board of the FCR into the system.

The last main part of the project will be the development of long-term back-up systems for the data and results of the different sub projects. In addition to the publication on print media, on CD-ROMs or on the internet the declarative modules of the overall system should be migrated into a format that is independent of the system – to the present knowledge it is XML – to be able to use also other hardware and / or software packages if necessary. For the procedural elements such strategies can not be foreseen yet and thus cannot be planned.

The main task of the applied project will be to unite all already existing sources and stocks at the same quality level, to develop them completely, to accumulate single sources and their developments for a whole solution and finally to export the overall system in a format based on XML.

2.4 Methodological considerations

Those parts of the project series of which the aim is to complete, correct and standardise already existing data of the virtual archive will be done according to the method of the 'integrated computer-supported edition (ICE)'. Controlling and evaluation will take place in the usual way after the 4 or 6 eyes principle (i.e. at least two people will cross-check). The innovative part of this is the possibility to integrate the expert knowledge of the scientific advisory board that will act as a peer-group prior to the final release of the material.

Another methodological innovation concerns the control of the processing steps about metadata in the virtual archive. Depending on the progress of work (job-specific) and on the source type (system-inherent) the existing material can be divided into different representation forms which affect themselves mutually. According to the type of the source there must be different basis- and documentation-modules. A meta data base in the virtual archive will provide the user with information about the existence and the state of progress of these modules. This is imperative because the user will obtain admission even to parts of the data base of the virtual archive which are not fully completed yet. In other words this means that the user has access to images, texts or documentations of sources and can use this material already before any edition of those sources are released. If accessing such preliminary material the user will be told about the 'state of aggregation' of the respective source by using the menu point of the respective web pages. The menu points of such web pages will display the processing step of the source.

The overall documentation of the entire stock will be achieved in two different ways. One is to accumulate the catalogues of persons and locations which are taken from each stock step-by-step in the virtual archive. This will automatically lead to a quality improvement in the area of the identifications of persons and locations for two reasons: (i) the source base will be enlarged, and (ii) the corrections made will also affect all different catalogues. In fact, those data bases which are used for compiling catalogues will be connected with the meta data base by complex record-linkage. This procedure is already methodically planned but not yet realized. The other way to obtain an overall documentation is to develop stringent thesaurus systems that are independent of a particular stock. By the administration of this declarative special knowledge in independent objects (the so-called codebooks) the original information of the source remains unaffected, i.e. no information gets lost. The user can access the source item as well as a terminology that is managed by a 'controlled vocabulary'. This makes it possible to represent the data in different ways and to classify them. The following example illustrates this action:

When analyzing account-books it is useful to classify all objects, services and intended purposes of both expenses and incomes by a thesaurus. To be able to demonstrate the objects' complexity a thesaurus with a five-digit code was compiled in the current project. Each digit stands for the respective sub-differentiation. After defining the code the appropriate code value must be assigned to every object mentioned in the source and in doing this classification modules are generated. Such modules are then available for further analyses that are independent of the database. Beyond this they can be used for retrieval operations and for the migration of data into formats which are 'understood' by statistic packages.

The already existing thesaurus systems illustrate external and internal features of a source according to the ICE-method. They will be extended to general codebook systems which are able to describe the

characteristic features of all source types in the entire stock. With the aid of such systems a stock-spanning 'subject index' will be realized which then can be used for statistical primary- and secondary-analyses.

Not only the revision of the existing data in all stocks and the development and extension of the virtual archive, but also the long-term survival of the overall system must be guaranteed. As already mentioned, migration of the databases with all features and ontologies into an independent format within the framework of the applied project is necessary. At present it appears that this format will be XML unless an important new development evolves in this area. In this context it will be necessary to develop appropriate data-type-definitions which are able to describe herein discussed data. In addition, routines must be developed which will manage the export from the FCR-system to XML thereby considering the definitions as outlined above. In the near future a pool of data will represent the core of the system in contrast to the actual set of software-depending objects.

References

Digital Preservation Coalition, http://www.dpconline.org/graphics/index.html, 2005-05-10.

Susanne Dobratz, Inka Tappenbeck, 'Thesen zur Zukunft der digitalen Langzeitarchivierung in Deutschland' in: *Bibliothek. Forschung und Praxis* 26/3 (2002), p. 257-261 [http://www.bibliotheksaur.de/2002_3/257-261.pdf].

Peter Doorn, 'Research data archives and public electronic record-offices: What can we learn from each other?' In: Doorn, Peter, Garskova, Irina and Tjalsma, Heiko (Eds.), *Archives in Cyberspace. Electronic records in east and west*, Moscow 2004, p. 96-111.

Hans-Heinrich Ebeling, Manfred Thaller (Eds.), *Digitale Archive. Die Erschließung und Digitalisierung des Stadtarchivs Duderstadt*, Göttingen 1999.

Lothar Gall, Rudolf Schieffer (Eds.), Quelleneditionen und kein Ende? Symposium der Monumenta Germaniae Historica und der Historischen Kommission bei der Bayerischen Akademie der Wissenschaften, München 22./23 Mai 1998 (Beihefte der Historischen Zeitschrift NF 28), München 1999.

Thomas Grotum, *Das digitale Archiv. Aufbau und Auswertung einer Datenbank zur Geschichte des Konzentrationslagers Auschwitz*. Frankfurt/Main – New York 2004.

Margret Hedstrom, 'Context and Custody: Strategies for long-term preservation of electronic records.' In: *Archives in Cyberspace. Electronic Records in east and west*, Moscow 2004, p. 129–148.

Rainer Hering, Udo Schäfer (Eds.), *Digitales Verwalten – Digitales Archivieren, (Veröffentlichungen aus dem Staatsarchiv der Freien und Hansestadt Hamburg 19)*, Hamburg 2004.l

Stuart Jenks, 'Die Verlässlichkeit von Informationen im Internet', in: *Internet-Handbuch Geschichte*, hg. von Stuart Jenks und Stephanie Marra, Köln-Weimar-Wien 2001, p. 265 – 271.

Roland Kamzelak, 'Hypermedia – Brauchen wir eine neue Editionswissenschaft?' In: *Computer-gestützte Text-Edition*, hrsg. v. Roland Kamzelak (editio Beiheft 12), Tübingen 1999, p. 119ff.

Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit (nestor), www.langzeitarchivierung .de, 2005-05-10.

Ingo H. Kropač, Heidrun Boshof, 'Digitale Edition eines umfangreichen Quellenkorpus, Vorgehensweise und Probleme bei der Aufbereitung, Strukturierung und Kategorisierung des Quellenmaterials' in: *Geschichte und Informatik – Histoire et Informatique* 11 (2000), p. 93-112.

Ingo H. Kropač, Ad fontes. *Von Wesen und Bedeutung der Integrierten Maschinellen Edition, in: Geschichtsforschung in Graz. Festschrift zum 125-Jahr-Jubiläum des Instituts für Geschichte der Karl-Franzens-Universität Graz*, hg. von Herwig Ebner, Horst Haselsteiner, Ingeborg Wiesflecker-Friedhuber, Graz 1990, p. 465-482.

Ingo H. Kropač (Ed.): *Fontes Civitatis Ratisponensis*, www.fcr-online.com, 2005-05-10.

Ingo H. Kropač, Susanne Kropač, 'Prolegomena zu einer städtischen Diplomatik des Spätmittelalters, das Beispiel Regensburg' in: *La diplomatique urbaine en Europe au moyen âge. Actes du congrès de la Commission internationale de Diplomatique, Gand, 25-29 août 1998* (Studies in Urban Social, Economic and Political History of the Medieval and Early Modern Low Countries 9), hg. von Walter Prevenier, Thèrés de Hemptinne, Louvain/Apeldoorn 2000, p. 229-265.

Andreas Metzing (ed.), *Digitale Archive – Ein neues Paradigma?* (Veröffentlichungen der Archivschule Marburg 31), Marburg 2000.

Record Management, www.nationalarchives.gov.uk/recordsmanagement, 2005-05-10.
Wilfried Reinighaus, 'Archive und Archivwesen', in: *Internet-Handbuch Geschichte*, hg. von Stuart Jenks und Stephanie Marra, Köln-Weimar-Wien 2001, p. 195 – 211.
Research Library Group (Ed.), *Trusted Digital Repositories: Attributes and Responsibilities*. An RLG-OCLC Report, May 2002 [http://www.rlg.org/en/pdfs/repositories.pdf, 2005-05--10].
Florian Rötzer, *Wider das digitale Vergessen*, www.heise.de/tp/r4/artikel/14/14211/1.html, 2005-05-10.
Seamus Ross, Edward Higgs (Eds.), *Electronic Information Resources and Historians: European Perspectives*. (Halbgraue Reihe zur Historischen Fachinformatik A/20) St. Katharinen 1993.
Patrick Sahle, 'Digitale Editionstechniken und historische Quellen' in: *Internet-Handbuch Geschichte*, hg. von Stuart Jenks und Stephanie Marra, Köln-Weimar-Wien 2001, p. 153 – 166.
Kevin Schürer, *Better access to electronic information for the citizen. The relationship between public administration and archives services concernung electronic documents and records management*, Luxemburg, 2001.
Michael Sperberg-McQueen, Claus Huitfeldt, Concurrent Document Hierarchies in MECS and SGML, in: Literary and Linguistic Computing 14 (1999), S. 29 – 42.
Sören Steding, 'Warum noch drucken?' in: *Jahrbuch für Computerphilologie* 3 (2001), p. 149-158.
Manfred Thaller, 'Ungefähre Exaktheit. Theoretische Grundlagen und praktische Möglichkeiten einer Formulierung historischer Quellen als Produkte ‚unscharfer' Systeme', in: *Neue Ansätze in der Geschichtswissenschaft*, hrsg. v. Herta Nagl-Docekal und Franz Wimmer (Conceptus-Studien 1), Wien 1984, p. 77-100.
Manfred Thaller (Hg.), *Codices Electronici Ecclesiae Coloniensis. Eine mittelalterliche Kathedralbibliothek in digitaler Form*, Göttingen, 2001.
Karsten Uhde, 'Urkunden im Internet – Neue Präsentationsformen alter Archivalien' in: *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde* 45 (1999), p. 441 – 464.
Gunter Vasold, 'External Features of Historical Documents in the Computer Supported Editing', in: *Congreso internacional sobre sistemas de información histórica. Ponencias y Mesas Redondas*, Vitoria-Gasteiz 1998, p. 173-177.
Colin Webb (Ed.), *Guidelines for the preservation of digital heritage*, unesdoc.unesco.org/images/0013/001300/130071e.pdf 2005-05-10.
Hartmut Weber, Gerald Maier (Eds.), *Digitale Archive und Bibliotheken. Neue Zugangsmöglichkeiten und Nutzungsqualitäten, (Werkhefte der Staatlichen Archivverwaltung Baden-Württemberg A/15)* Stuttgart 2000.
Michael Wettengel (Ed.), *Digitale Herausforderungen für Archive. 3. Tagung des Arbeitskreises 'Archivierung von Unterlagen aus digitalen Systemen'*, (Materialien aus dem Bundesarchiv 7) Koblenz 1999.

Collaboration on medieval charters – Wikipedia in the humanities?

*M.A.B. Burkard**

The online-encyclopaedia Wikipedia celebrated its fifth anniversary on January 15 2005. By now, it holds 1.4 million articles in more than 200 languages. The English version alone includes more than 500.000 entries (as of March 2005, Heise 2005).[1] The success of Wikipedia shows the great potential of collaborative working modes that are supported through a suitable online platform. Taken this background, it seems to be an obvious idea to also apply this approach to the field of the humanities. As within research a much stricter control of the contributions' quality is needed, however, a direct transfer of a 'wiki' would not suffice.

The aim of this paper is to describe the prototype of an editorial system for the collaborative editing of medieval charters. The prototype focuses on the creation of full transcriptions of these charters, which can additionally be marked up with semantic markup. While the first section briefly outlines the context of application of the prototype, part two describes its functionality in its present state. After the architecture of the software is outlined in part three, the fourth part shows a number of further developments that are necessary for a production version of the editorial system. It has to be emphasized that this is a work-in-progress report that is not describing a fully developed product.

The context

The specific context of for the development of the prototype is the MOM-Project[2]. The MOM-Project, which is organized by the archive of the diocese in St. Pölten, Austria, is aiming at digitizing all medieval and early modern charters of Lower Austria's monasteries. Also, these images are to be supplied with descriptive metadata including regests or full transcriptions. This way, approximately 20.000 charters will be made available on the web until mid 2005.[3] It is planned that the developed prototype 'EditMOM' will be developed further into a production version and integrated with the MOM-Project as an additional module.

The prototype – a 'walk' through the system

To start working with the system a user has to register onto the system by providing some personal information and a valid email-address. The user can then navigate through the holdings of the MOM-Project by using a browsing component and view the digitizations of the charters along with the descriptive metadata. In case a charter does not have a full transcription associated with it, the user can create one. Besides typing in the full-text the user can add extra information by tagging the text with semantic markup.

To work on these tasks, the user is provided with an editorial environment that consists of a frameset which is horizontally divided into three parts. In the upper frame the first scan of the charter is shown. As it is scaled horizontally to fit exactly the width of the screen, horizontal scrolling while reading the charter becomes unnecessary. If parts of the charter are difficult to make out it is possible, however, to show the scan in its original resolution and thus zoom in to the desired details. In the middle frame, links to further scans can be found in case a charter consists of more than one page. Finally, the bottom frame displays an editor for creating the transcription. The editor consists of a text field for typing in the transcription as well as a menu bar and some toolbars for controlling

* University of Cologne, Cologne

1 The English 'Wikipedia' version can be found at: http://en.wikipedia.org/wiki/Main_Page.

2 'MOM' is the common abbreviation used in medieval charters for the Latin word 'monasterium'.

3 Cf. http://www.monasterium.net.

the functions of the editor. One important feature of the editor is that it supports the user in tagging the full text of a charter with a defined set of XML-elements.

This tagging of the full texts of charters has a number of advantages, one of which being the support for advanced retrieval functions. The prototype offers, for instance, the option to tag persons in the text and to associate them with the appropriate entry from a controlled vocabulary. This way documents in which certain persons appear can not only be found (as is already possible today on the regular MOM-Project pages), but it is also possible to jump directly to the exact passages within the text.

Another possibility of an improved retrieval results from the tagging of the typical structure of charters (*formula*)[4], which is also supported by the prototype. As different branches of research mainly come to their findings from studying only certain parts of the *formula* – Economic History, for instance, is interested in the *narratio* and the *dispositio*[5] – functions that limit the retrieval to the desired parts of the charter can be of great help to researchers.

XML-tagging can also be used for creating different layers of a text. Printed charter editions normally present normalized versions of the texts[6] which facilitate a content-orientated, fast reading. For some groups of researchers, however, the focus in dealing with charters is not on their readability. For historic linguists, for example, an edition should be as close as possible to the original source. With the help of XML different text layers can be created. For instance, one layer can display a charter as authentic as possible while at the same time a normalized version is provided so that different user interests are satisfied. To exemplify this, the prototype offers an option to resolve abbreviations (which are then carried as an XML-attribute), while at the same time the 'original' text layer is preserved containing the exact spelling of the charter.

Additionally, it is possible to make annotations that have to be set at a certain cursor position. The possibility of tagging whole text passages has not been provided in the prototype as this could easily lead to the problem of overlapping hierarchies that can not easily be dealt with in XML. There are suggestions for solving this problem (Czmiel 2004) but their integration into the prototype would have exceeded the scope of this project.

The tagging procedure is otherwise very similar for every element. First of all, the user marks the desired text passage, and afterwards presses the button for the corresponding XML-tag. If it is optional to add an attribute to the respective element, a text input window opens for typing in free text (*e.g.* for resolving an abbreviation), or the user is provided with a controlled vocabulary for the respective index (persons, objects, places etc.) from which the user can choose the appropriate term. The program adds the XML-tags internally at the chosen position, and afterwards validates the XML-document against a simple DTD that was written for marking up the charters within 'EditMOM'. If an error occurs during the validation process, the XML-tag is not inserted into the text field. Also, the user receives a message telling him which mistake he has made. Otherwise, the XML-tags are included into the text and the user can continue editing the charter.

Another function of the prototype is that the user can switch between the edit mode and different view modes. Clicking on one of several view-buttons causes HTML-views to be generated on the server and sent to the client. For instance, it is possible to view all index terms which are then graphically highlighted and provided with a tool tip holding information about the respective index type and the ascribed term from the controlled vocabulary. Other views show the 'original' text layer as well as a normalized version with resolved abbreviations or display the *formula*-parts of the charter in different colours (also provided with tooltips).

At least with extensive charters the time needed for editing can be quite long. Therefore, the user is permitted to work on a certain charter and 'lock' it for his own use for a longer, though limited period of time, thus making sure that two users do not introduce inconsistent contributions. A storing mechanism is provided that allows the storage of a likewise limited number of documents. Both limitations are to prevent

4 The *formula* is a general scheme of charters that has been distilled from a huge number of medieval charters.
For the terminology see http://pcghw51.geschichte.uni-muenchen.de/UrkDTD/vid.php.

5 While the *narratio* describes the factual or alleged circumstances of and causes for the issuing of the charter, the *dispositio* names the actual legal act or the core of the charter (Brandt 1998, p. 91).

6 In normalized texts abbreviations are normally resolved, punctuation and capitalisation are adapted to modern conventions etc.

the 'hoarding' of greater numbers of documents which then could not be worked on by other users.

Ensuring the high quality of the collaborators' contributions is of great importance, as the charters of the MOM-Project are used as the basis of scientific research. It is to be expected that, when collaborative working modes are employed, acceptance by the scientific community of medievalists can only be secured if scientific standards are met. In order to guarantee the quality of the collaboratively edited transcriptions 'EditMOM' includes the function of a moderator. When a user finishes editing a charter he unlocks it. A moderator then gains access to all charters that have been unlocked. He corrects the contributions if necessary and finally decides whether they should be transferred from the storage database to the regular database or rejected and deleted. The qualification of a moderator has to be beyond doubt and this person should probably be a member of the MOM-Project. The existence of such a highly qualified authority seems essential to ensure high scientific standards.

The architecture of the software

The architecture of the system is made up of three parts that interact with one another. The backend consists of a database which holds all the data. The server functions as middleware between the database and the client. It passes requests made by the client on to the database and sends the answers back to the client. With 'eXist'[7] the prototype 'EditMOM' uses a native XML-database that is entirely programmed in Java and among others accessible via XPath, XQuery as well as the XML:DB-API. The server functionality is implemented as Java web application in which the Struts-framework is used. A common web browser with Java Plug-in and a Java-Applet acts as the client component; the latter serving as the editor for transcribing the charters and adding the semantic markup to them.

The future

As mentioned above, some further developments are necessary for a fully functioning production version. Naturally, there are hardly any limits for enhancing such a system, especially with regard to deeper semantic and structural markup that could be added to the charters. Therefore, in the following only the most relevant aspects are named.

a) 'Hiding' of the markup: In the present developmental stage of the prototype the XML-tags are visible for the user. The goal is to hide the actual markup from the user and only make it visible through visualisation techniques. On the one hand, this can help to avoid mistakes during the tagging process. On the other hand, it can also raise the level of the tool's acceptance with those scientists who are not very familiar with the new media. A mode for 'hiding' the tags has unfortunately not made it past an advanced test stage, yet it should not be missing in a production version.

b) Extension of the role concept to relieve the moderators: In the present stage of development of the prototype the role of the moderator is extremely important for guaranteeing high scientific standards. It is obvious that this part of the collaborative work process could easily become a bottleneck due to a shortage of personnel within the project. Therefore it seems important to find ways that can help to reduce the need for professional personnel, even though it is hardly possible to cut it back to zero.

An approach to be pursued for the production version of 'EditMOM' is to assign graded rights to different user groups. This aims at harmonizing the degree of complexity of an 'allowed' task and the degree of qualification of the editor. For instance, every user who is registered onto the system should be allowed to create transcriptions. More complex tasks should only be performed by editors whose qualification in the subject is out of question (*e.g.* because of personal acquaintance with those in charge of the project) or who have proved their qualification through the completion of 'easier' tasks. The qualification for extended rights could be organized through an assessment system in which the moderator evaluates each contribution. After obtaining a fixed number of positive assessments the user could be allocated a new role with additional rights.

In order to further reduce the work load for those responsible for the project, especially qualified editors could be trusted with tasks of a moderator and assigned with an appropriate role. These 'co-moderators' could then be in charge of correcting the contributions of regular users and thus relieve the personnel

7 Cf. http://www.exist-db.org.

of the MOM-Project so that ideally these only have to exercise an overall supervision.

c) Versioning system: Currently it is only possible to create *new* transcriptions. A versioning system is needed if users should also be able to correct or complete (*e.g.* with additional markup) *existing* data as otherwise the moderator would not be able to effectively correct the changes made.

On top of that a versioning system would allow for truly collaborative work. The idea resembles the concept of Wikipedia and is basically as follows: A user starts working on a transcription. Once, he has finished it, a recent changes function tells other users about this new entry and the community can apply changes to the document – each saved as an own version – for a certain period of time until a moderator finally decides whether the transcription is transferred to the regular database. This way, the moderators should be relieved of at least some of the correction work as part of it would be taken over by regular users.

To make this possible, the editorial system should not only be capable of retaining the different versions but should additionally provide a function that allows easy comparing of the different versions. When two versions are compared with one another, all changes should be clearly highlighted so that the moderator can easily find the respective modifications. For each version it should furthermore be recorded who created it so that the person in question can be contacted by a moderator or another user if necessary. To promote communication between the collaborators it would also be helpful if comments could be added to changes. This way, a user could explain the reasons for the changes they made so that they would be made more comprehensible for a moderator or other users.

d) Using standards: It seems sensible to pay attention to the efforts of the Charters Encoding Initiative (CEI) which, following the TEI, aims at developing a general standard for the tagging of charters and has already suggested a corresponding XML schema (Vogeler 2004).[8] Unfortunately, the author heard about these efforts too late to follow the CEI-schema inside the prototype so that a self developed DTD is used for the system presented here. However, the CEI-schema should at the latest be used in case it is adopted as an official standard by the TEI since this would make the integration with other editions on the web much easier.

In the end, one has to wait and see in how far scientific communities will make use of the possibilities of collaborative working modes as can be offered by systems like the presented prototype 'EditMOM'. Maybe it is possible to repeat the success of Wikipedia in the context of the humanities. Providing software that meets the specific requirements of the particular scientific community is in any case an indispensable prerequisite.

References

Brandt, A., von (1998). *Werkzeug des Historikers. Eine Einführung in die historischen Hilfswissenschaften*, 15. Aufl., Stuttgart – Berlin – Köln.

Czmiel, A. (2004). *Adäquate Markupsysteme für die digitale Behandlung altägyptischer Texte.*
URL 21.03.05: www.hki.uni-koeln.de/studium/MA/MA_czmiel.pdf

Heise-Newsticker (2005). *Englischsprachige Wikipedia feiert 500.000sten Artikel*.
URL 21.03.05: http://www.heise.de/newsticker/result.xhtml?url=/newsticker/meldung/57729&words=Wikipedia

Vogeler, G. (2004). 'Ein Standard für die Digitalisierung mittelalterlicher Urkunden. Bericht zum Workshop (München 5./6. April 2004)', *AHF-Informationen* 23.
URL 21.03.05: http://www.forschung.historicum.net/tagungsberichte/muenchen-200404.html

8 The schema can be found at: http://pcghw51.geschichte.uni-muenchen.de/UrkDTD/cei.xsd or a graphic overview of the schema can be viewed under at: http://pcghw51.geschichte.uni-muenchen.de/UrkDTD/Schema-Overview.html.

Reinventing the humanities in a networked environment: The Australian Network for Early European Research

*Toby Burrows**

The Network for Early European Research (NEER) is a new Australian initiative to broaden and deepen research in the field of medieval and early modern European studies. An integral part of the Network is a Digital Services Programme, which integrates various digital strategies designed to advance the Network's ambitious goals. This paper looks at the ways in which the Network is using information technologies to reinvent a traditional humanities discipline.

Research networks, cyberinfrastructure and the humanities

Academic research is being changed fundamentally by two major imperatives. In the first place, it operates increasingly within an interdisciplinary and international framework. Research teams are increasingly international, and are tending more and more to be composed of temporary groupings of researchers from a range of disciplines, brought together to address a specific problem. The complexity and scale of these research problems require the assembling of expertise from different disciplinary perspectives as well as from different institutions and organizations. Research into the problem of soil salinity, for example, requires the involvement of soil scientists, environmental scientists, plant biologists, animal biologists, water engineers, and even historians.

Closely allied to this is a second trend: the growing centrality of information technologies to the ways in which research is done. The importance of information technologies in academic research has been highlighted by a number of recent initiatives and investigations. The Atkins Report to the US National Science Foundation in 2003 used the term 'cyberinfrastructure' to describe the way in which software platforms can be interlinked to manage the vast deluge of scientific data, enabling researchers around the world to tap into an international grid of digital research (Atkins 2003). In the United Kingdom, e-Science and e-Social Science programmes have been established to channel funding into the development of such infrastructures (Hey 2003). In Australia, the term e-Research has been used for the same type of initiative.

These trends are already evident in disciplines such as astronomy, where national and international groupings of researchers are being formed through the agency of 'virtual observatories' (Djorgovski 2004). By linking together shared data from observatories around the world and providing appropriate software tools to analyse and manipulate the data, astronomers are increasingly able to work together across national boundaries as part of a single global knowledge community.

This kind of approach is well-suited to scientific research, but its appropriateness for the humanities is yet to be tested. Of particular relevance and interest, then, is the Commission on Cyberinfrastructure for the Humanities and Social Sciences, established by the American Council of Learned Societies in 2004. The Commission is intended to go beyond the use of digital technologies as 'tools enhancing research methodologies', and to examine ways in which they can become 'a force creating environments that enable the creation of new knowledge' (ACLS 2004). A range of experts in the humanities and social sciences have been presenting the Commission with their ideas for the creative and innovative use of technology to transform their disciplines.

In Australia, the main government funding body for research in higher education institutions – the Australian Research Council (ARC) – has responded to these imperatives by establishing its Research Networks programme in 2004. The purpose of this programme is to build large-scale groups of researchers and encourage

*Digital Services Director, Network for Early European Research, University of Western Australia

them to collaborate across institutional and disciplinary boundaries. It goes beyond the ARC's existing support for smaller-scale collaboration, and aims to develop these linkages at a national and international level. The ARC has also initiated an e-Research programme, but this is as yet on a fairly small scale. The funding for each Research Network is of the order of A$1.5 million over a period of five years. Twenty-four networks were funded in 2004, mostly in the areas of science, technology and medicine.

One of the key areas in which the ARC envisaged that the networks would be active was the development of shared information technologies and knowledge management tools, new databases, and new technologies for communication and interaction. These activities were seen as part of the crucial infrastructure which would be needed to underpin collaborative research in a national setting.

The Network for Early European Research

The Network for Early European Research (NEER) was one of only two ARC networks to be funded in the humanities. It is officially based at the University of Western Australia, where its executive and secretariat are located, but most of its academic activities (conferences, seminars and workshops) take place 3,000 kilometres away on the Eastern side of Australia. The Network's structure is a mixture of individual researchers and institutional members. Its individual participants include researchers in most of Australia's 37 universities, ranging from eminent academics through to postgraduate students and early career researchers. More than 150 individuals are currently listed as Network participants. Their research covers all aspects of the culture and history of Europe in the Middle Ages and the Early Modern period, extending up to the initial European connections with Australia in the late eighteenth and early nineteenth centuries.

The Network has a range of institutional members including most of the larger Australian universities, such as Melbourne, Queensland and Sydney, which are all making a financial contribution to the Network. There are also a number of institutional partners, including commercial publishers like Brepols ProQuest and the University of Western Australia Press; public collecting institutions like the State Libraries of New South Wales and Victoria, and the Western Australian Maritime Museum; and community groups like the Perth Medieval and Renaissance Group, Australians Studying Abroad, and the Woodside Valley Foundation.

The Network is organized around four main themes or research problems: Cultural Memory; Social Fabric; Science, Medicine and the Environment; and Early European/Australasian Connections. Each of these has a team leader, whose role is to coordinate Network activities and communication between researchers with an interest in the specific research area. Most of the academic activities of the Network are focused around these four themes, with each area organizing and supporting conferences, seminars, postgraduate advanced training workshops, and meetings to develop a research agenda and develop collaborative grant applications.

While most of these are the type of activity in which academic researchers have been regularly involved, the difference in the Network's case is that they take place at a national level and within a comprehensive national framework. They are not simply reliant on university departments, learned societies or even individual researchers. The Network contributes financially, and also has an overall strategy for encouraging participation by postgraduate students and early career researchers. The other important difference is that these are only one strand in the Network's programmes. Its activities in the digital arena are where it aims to be most innovative and forward-looking.

Digital Early European Resources: Stage 1

In response to the ARC's emphasis on the use of information technologies within its Research Networks, NEER has developed its own digital services agenda as an integral part of its activities. Known as Digital Early European Resources (DEER), this programme brings together various activities in the digital arena which have two main goals: to provide resources for the Network's participants, and to enable them to communicate more effectively with each other.

The Network's initiatives in promoting more effective communication among its participants are focused partly on such well-established mechanisms as mailing lists, discussion groups, and a database of participants. But other activities are more innovative: the Network has set up a scheme known as e-Consult, which enables postgraduate students to identify and contact senior researchers who are willing to provide advice. While the initial contact is made through the Network's Web site, and includes an agreement to

observe the protocols set down by the Network, subsequent communication is entirely confidential to the parties concerned.

Providing resources will be done in two main ways. The Network is working in partnership with two major commercial publishers of specialist databases to provide access for Network participants whose institution does not have a subscription. ProQuest is providing access to Early English Books Online (EEBO) while Brepols is providing access to a selected number of its full-text products. The main beneficiaries of this approach are researchers in smaller universities and regional universities.

At the same time, NEER is developing its own resources, beginning with a discovery service for identifying Early European artefacts, artworks and manuscripts in Australian collections. This involves a federated search service across the records of different types of institution with relevant collections: libraries, museums, archives, and galleries. Where digital versions of these objects are already available, the Network will enable researchers to find and view them.

As far as possible, the Network aims to avoid duplicating work already done in the cultural heritage institutions themselves, and prefers to harvest existing metadata from their sites. The same principle applies to the digital objects identified through the resource discovery service. The Network prefers to point to a file on the server of the appropriate institution, and avoids creating or storing its own digital files. It is, however, sponsoring a digitization programme on a small scale, by identifying significant objects which have not been digitized and working with holding institutions to digitize them. Among the models for this resource discovery service are PictureAustralia and MusicAustralia, both of which are produced by the National Library of Australia (Ayres 2004).

The Network is also involved in electronic journal publication. The refereed journal *Parergon*, which is published by the Australian and New Zealand Association of Medieval and Early Modern Studies (ANZAMEMS), is issued in electronic form through the Network's Web site, as well as being available through journal packages like Project Muse. NEER provides publishing and subscription management services for this electronic version of *Parergon*.

Another area in which the Network has a keen interest is the training of postgraduate students and early career researchers in specialist skills relevant to Early European research. The Network's digital programme is contributing to this goal by establishing internships in partnership with publishers of electronic text collections. Postgraduate students are given training in text encoding with markup languages like TEI as well as in database management and the development of Web sites. This training is closely linked to the use of editorial and research skills to create and review specialist content for databases in the fields of medieval and Early Modern history and literature.

Digital Early European Resources: Stage 2

While these initial activities in the Network's digital services programme are integral to its success, they are not designed to be technically innovative, nor are they expected to promote transformative cultural change in humanities research. They are intended to harness and integrate a wide range of proven uses of technology. The next stage in the Network's development, however, is designed to use information technologies to reinvent and transform the discipline in a more innovative and fundamental way.

One of NEER's main priorities in this second stage is to develop its own digital repository for the research output of its participants. To date, most digital repositories of this kind have been institutional – that is, they reflect the output of a single university (Lynch 2003). While there has been considerable support and enthusiasm at the level of university policy and in university libraries for establishing these repositories, it has to be said that this approach has not been particularly successful in practice. Most institutions have had difficulty acquiring a sufficient body of material for their repository.

NEER is working with the ARROW Consortium (Australian Repositories Online to the World), based at Monash University in Melbourne, to develop a repository which is discipline-based and national in scope. ARROW has been funded by the Australian Government as part of its Research Information Infrastructure Framework for Australian Higher Education. Its goal is to identify and test software to support best practice institutional digital repositories (Payne 2005). ARROW is using the Fedora architecture and software, initially to manage such materials as article preprints and postprints, theses and other print equivalents. The NEER repository will not be limited to articles and papers, though. It will also aim to collect and make available the underlying research data from these stud-

ies, whether in the form of databases, spreadsheets, correspondence, images, sound files, maps, or other formats. NEER's expectation is that researchers will see the value of contributing their research output to a repository as part of their continuing communication with fellow researchers in their discipline. By embedding the repository into the existing pattern of disciplinary communication, the Network aims to provide a sufficient incentive for researchers to participate, and to encourage the cultural change which is required for such approaches to be successful.

Within this repository, the Network is focusing particularly on metadata content and standards. While various institutions and projects, including ARROW, have worked on metadata schemas and on the mapping of data from one schema to another, comparatively little work has been done on the content and standards for repository metadata. Most of the subject access to repositories is at a fairly simple and summary level, and researchers depositing their papers are given only a limited choice of subject terms for describing the content. NEER is investigating methods for embedding sophisticated subject ontologies into the repository framework, by bringing together the specialist disciplinary knowledge of the participants in the Network and existing vocabularies for subject access in indexing databases like the *International Medieval Bibliography*.

Early European research is a difficult area for metadata because of the many European languages used in the original sources and in contemporary scholarship, and the lack of consistent terminology in some fields. Mapping variant forms of names is a particular challenge. As far as possible, the Network intends that this work on metadata and subject vocabularies should link into the broader framework of the Semantic Web.

The Network is also very interested in exploring how the existing repository framework might be used in transforming the way in which research is communicated and disseminated. At present, repositories are a mainly static service. They collect and archive research papers and articles and make them available to other researchers. This underlying goal is not significantly different from that of research libraries, though the communication system on which they are based does not depend so much on publishers and journals. The immediacy and availability of research results are improved, but the effect of repositories on the research process itself is not transformative.

NEER is aiming to test the integration of repositories with other emerging technologies in an effort to design new structures for communicating research in the humanities. It aims to make room for younger researchers especially to experiment with new approaches. Among the components of this structure will be blog-like narratives of research activities, seen as a continual work in progress, but designed to report research results. These will link to research data and source materials, housed in databases and similar structures assembled either by the research group or externally by other groups, including publishers. They will also link to more formal publication of results in such structures as repositories, journals and monographs. This framework will be a record of both individual and collective activities within the national Early European research group.

This kind of structure will open the workings of humanities research to a more continuing scrutiny than is possible in the traditional system of publication. It will also promote this research to a much wider audience across the Internet, through exposure to Google Scholar and interlinking to external subject gateways and similar sites. A crucial element will be the incorporation of methods for enabling and recording peer review of the research. Without this, any new structures for disseminating research are unlikely to rise much above vanity publishing. Providing avenues for other researchers to evaluate, comment on and respond to research will be critically important in a truly transformative use of the digital environment.

In these ways, NEER is aiming to contribute to the design of 'a next generation system for scholarly communication' (Van de Sompel 2004) which goes beyond Open Access, self-archiving and electronic journals. Such a system will need to redefine the 'unit of communication' and provide a more flexible way of 'registering' a communication unit. The Network is able to work closely with a vigorous national community of researchers to design and test new ways of distributing and evaluating their communications.

Conclusion

The main goal of the Network's digital strategy is to promote collaboration and communication between humanities researchers at a national, discipline-based level. This is being done by harnessing and integrating a wide range of information technologies. In the initial stage of the Network, these are well-established and widely used already, but others are new approaches

which are intended to transform the way in which research is communicated in the humanities.

While the activities of the Network include software development and technical innovation, this is not the primary focus. Nor is the Network emphasizing technical solutions to the management and manipulation of large datasets along the lines of the British e-Science programme. Instead, the Network's main emphasis is on transforming the way in which researchers communicate within a well-established humanities discipline, and on fostering the kind of cultural change which will be a necessary part of this transformation.

NEER offers a unique opportunity to develop and test new technologies for scholarly communication and the dissemination of research, within the framework of a new government approach to funding academic research communities. This digital programme will be crucial to the Network's success in reinventing Early European research and demonstrating the applicability of new technologies to fundamental changes in the humanities.

References

ACLS. 2004. Commission on Cyberinfrastructure for the Humanities & Social Sciences. *Charge to the Commission*. New York: American Council of Learned Societies. http://www.acls.org/cyberinfrastructure/cyber_charge.htm

Atkins, Daniel E. et al. 2003. *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the Blue-Ribbon Panel*. Arlington, VA: National Science Foundation. http://www.cise.nsf.gov/sci/reports/toc.cfm

Ayres, Marie-Louise, Toby Burrows and Robyn Holmes. 2004. 'Sound footings: building a national digital library of Australian music' in: *Research and Advanced Technology in Digital Libraries,* ed. Rachel Heery and Liz Lyons *(Lecture Notes in Computer Science 3232)* (Berlin: Springer-Verlag, 2004), 282-291.

Djorgovski, S. G. 2004. 'Virtual Observatory, Cyber-Science and the Rebirth of Libraries', ARL/CNI Forum on E-Research and Cyberinfrastructure. http://www.arl.org/forum04/djorgovski.html

Hey, A. J. G. and A. E. Trefethen. 2003. 'The Data Deluge: An e-Science Perspective', in: *Grid Computing – Making the Global Infrastructure a Reality*, ed. F. Berman, G. C. Fox, and A. J. G. Hey. (London: Wiley, 2003), pp. 809-824.

Lynch, Clifford A. 2003. 'Institutional Repositories: Essential Infrastructure for the Scholarship in the Digital Age', *ARL Bimonthly Report* 226. http://www.arl.org/newsltr/22/ir.html

Payne, G. J. 2005. 'Australian Research Repositories Online to the World – ARROW', EDUCAUSE Australasia, Auckland, April 2005.

Van de Sompel, Herbert, et al. 2004. 'Rethinking Scholarly Communication: Building the System that Scholars Deserve', *D-Lib Magazine* 10 (9), Sept. 2004.

Digitising parish registers – principles and methods

*Nanna Floor Clausen**

Introduction

Since 1992 a project has been going on with the aim of creating a machine-readable version of the Danish census records. The project has during this >10 years period gained much experience and has become widely known. The reason for initiating the project was the identification of a large group of people already transcribing sources in order to take a copy of them back home from the archive. The transcriptions created in this way were very disparate in quality and the format used depended on the knowledge and software at hand for each person. But what was even worse was that this transcription was normally only known by the person who had made the transcription. The result was that many people could be transcribing the same sources over and over in different places at different times.

A group consisting of experts representing professional historians and the genealogical organisations established themselves as the Source Entry Project. The focus for the project group was to define a format for the sources and to make this format widely accepted. The next focus was to find a way to organise the transcription of the sources in such a way that a source was never transcribed more than once. Having the Danish Data Archive organise, archive and distribute the transcribed sources solved the last problem.

The Source Entry Project had ambitious goals and set out to define formats for both structured and unstructured sources. Structured sources are sources where you have the same structure in the source for the same period and/or geography. Structured sources are sources like censuses, some parish registers and land registers. Unstructured sources are sources with no structure like tenancy records and manorial court rolls. The older parish registers constitute a group in between you can define as semi-structured. Some of these parish registers are written in a very irregular manner and it is difficult to define a structure whereas others from the same period but kept by another vicar are very structured and accurately kept. The ambitions of the group were also reflected in the defining of two sets of formats for the structured sources: a basic model and an advanced model. The advanced model imposed very high demands on the persons using this model as *e.g.* a name should be split in three parts: first name, patronymic and rest of the name. And the name as in the source should also be entered. The first name and the patronymic should be standardised so names considered as the same name but spelled differently should be standardised to the same name.

The date of birth should also be entered in three fields: date as in the source, date in a normalised text form and finally date in a database standard format.

The requirements for using the advanced format were so big and the debates on *e.g.* the standardisation of names were many so in reality the advanced formats have not been used for many years.

From the beginning it was not the intention of the group to develop a program for transcribing sources but volunteers developed a program that could handle the formats defined. This was a great achievement for getting the transcribed sources in a uniform format. Since the beginning of the transcription really began to gain momentum from 1994 a new program has been developed and stricter guidelines have been developed. The quality and specification of the formats for censuses and parish registers respectively has had the effect that focus has been almost solely on censuses. The good thing being that several censuses are now completed for Denmark and more censuses will be completed within the next few years.

Earlier efforts to transcribe the parish registers

The parish registers are as important as the censuses for historical demography. For genealogists they are equally important and they are pressing for transcribing parish registers. The formats defined for parish

*Danish Data Archive

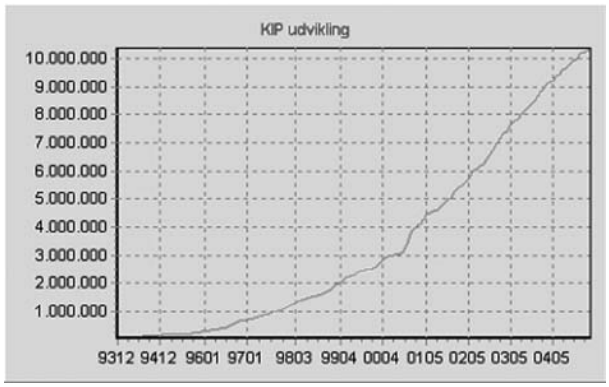


Figure 1. The development in number of transcribed records since the Source Entry Project began

registers were not as good as they should be. One major reason for this is the variety of structures of the parish registers and the lack of structure for many of them. Despite this an attempt was made to develop a program that could deal with the transcription of the parish registers. Though this program was developed in 1994 and using DOS it is still the only program existing. But it is practically not in use any more.

A few years ago a new attempt was made to develop a program for transcribing the sources. The DOS program developed in 1993 – 94 was based on the formats defined in the Source Entry Project group. This format was reasonably well described and reasonably close to the sources but nevertheless it was difficult to use. The amount of described fields was 175 which may seem as a lot but in fact it turned out to be neither enough nor correctly enough defined.

This lead to some frustration and as mentioned a new attempt to develop a more comprehensive and precise definition of the format was made a few years ago. The work was based on the previous format with amendments. The ambition of this program was very high as it intended to include every bit of information from all types of events in the parish registers. The result turned out to be too complex and too difficult to use. The format was made in a way that took all possible variations of the parish registers into account in order to make a complete transcription of the parish registers. The output was XML files.

The result of this situation – with no commonly accepted formats and not any program to use for the transcription – is that the volunteers and genealogists have started either defining their own formats and to use many different kinds of software or are pressing hard to get a common standard for parish registers.

The parish registers – their history

The chronology of the parish registers can be roughly divided into three periods: the first one being prior to 1812/1814, the second from 1814 – 1890 and the third from 1891 and onward. As the source material differs totally within these periods they will be treated separately.

Prior to 1812/1814

In 1645 and 1646 royal missives were sent to the bishops. These missives prescribed how the vicars should 'keep correct parish registers having dates for children born, how many are married and how many died'. These missives were followed to a large extent and many of the parish registers are kept from this period and onward. Before 1645 only very few parish registers had been kept and still fewer are preserved. The number is too little to be useful for demographic analysis. During the period 1645 – 1814 the quality of the parish registers improved and the more the closer the time is to 1814. Quality in this contexts refers to more detailed information on persons, i.e. their own names in stead of referring to a person as 'NN's wife', ages were added to the registration of people who had died and often more events were written into the parish registers than earlier.

Though the bishops inspected the vicars' work with the parish registers with respect to their content and level of information the structure of the parish registers differed widely over time and place. In the central laws it was stated what kind of information the vicars should keep but the way it was done depended upon the vicar. The majority of the earliest parish registers are generally referred to as being of the type 'chronological'. A vicar kept a register in which he entered baptism, death etc. in the order in which they appeared or in the order in which he remembered to enter them.

1812/1814 to 1890

In 1812 a new decree was sent out which completely changed the way in which parish registers were to be kept. In the decree were detailed regulations of what kind of events should be entered and how it should be done. This was done by developing ledgers with preprinted forms for the following events: 1) male baptisms, 2) female baptisms, 3) male confirmations, 4) female confirmations, 5) marriages, 6) male burials,

7) female burials, 8) in-migrants, 9) out-migrants, and 10) an alphabetical index for all events making it easy to find each entry. The decree also demanded that the parish register should be kept in two copies, one by the vicar and one by the parish clerk and that the copies must never be kept 'any night under the same roof'. Twice a year the vicar and the clerk should compare their entries to ensure that their contents were identical. This change in the registration procedure was carried into effect during 1813 and 1814 in most parishes and had as a consequence not only that the geographical coverage now became universal but also that the information became of higher quality and much more homogeneous.

The use of the pre-printed forms secured that the various other types of information demanded were collected by nearly all vicars. At the same time a few vicars continued entering additional personal data under the various headings or in the field 'comment'.

The parish registers from the period after 1814 are referred to by their type of event.

1890 and onwards

The changes in the parish registers were minor compared to the reform from 1812. New forms for the reporting of all demographic events were introduced. Much more detailed information on bride and bridegroom concerning their birth dates and birthplaces, civil status and also about the parents of the spouses. Place of birth was introduced in burial registrations. These forms have been in use until electronic registers replaced them a few years ago.

General methodology

Transcription of sources has implications for the use of the sources and on the information you can get from the sources. When converting a source from the original paper form to an electronic format you loose information on some areas. The obvious one is the 'look and feel' of the source – the physical representation. If this is considered important for the use of the source the user will have to use the source itself and not the transcribed version. When transcribing the source you will have to make decisions on the interpretation of the handwriting and get an understanding of what is actually written. The reading process is already an interpretation of the source where decisions are made more or less explicitly. The next very important interpretation is the actual process of creating the machine-readable version of the source. The crucial decisions are how and how much shall be transcribed: a transcription as close to the source as possible, a formatted transcription where the information is transcribed as *e.g.* fields, all the information is entered as text or only a subset of the source is registered? The answers to these questions are depending on the way you imagine the use of the source. Is the transcription going to be used in a special project just by yourself, shall it be used for statistical purposes, shall it be used by an unknown user group with all kinds of interests in the source? No matter which solution is chosen the solution must be documented. Finally the storage of the transcribed data has importance for the present and future use.

For the transcription of the parish sources these questions were of importance and required answers. As described above the parish sources can be divided into two types: a structured format after 1814 and an unstructured – or rather semi-structured – before 1814 (with many local exceptions). This structure (and lack of structure) together with the experience from the transcription of the historical censuses led to the decision of using a structured format when transcribing the parish registers. The censuses are preserved in a relational database and this storage format was also decided for the parish registers. It was never a question whether or not the transcribed material should be preserved at a central place or not. It was an assumption from the beginning that the data should be preserved at the same place as the material from the census project. This will furthermore benefit the usage of the parish registers as they may in this way be used for research in combination with the censuses. And vice versa enhance the value of the transcribed censuses.

The next answer to the questions above concerned the use of the transcribed parish registers. The experience from the census project is that the usage is not restricted to a narrow user group. Both historical demographers and professional historians use the material as well as genealogists and amateur historians. The result of the transcriptions must consequently be organised in a way that allows for the requirements from these identified user groups. One important aspect is here that it must be stated in the transcribed data from which source the transcription was done. In this way a user can make a check of the transcription against the source. A transcription is always an interpretation of the source and therefore the precise reference to the original source is essential.

The difficulties experienced in defining formats for

transcribing the parish registers referred to above were taken into account when defining a new format for the parish registers. It was early in the process very clear that the goal was not a complete transcription but instead the goal should be a 'register' or an 'index' to the original source. This decision is built on the logic/understanding that a) a transcription is anyway an interpretation of the source, b) the parish registers have a very complex structure varying over time and finally c) that it is too large a task to make a complete transcription of a parish register. Having taken this decision an analysis of what information was essential and of what could be left out was necessary.

Principles for the index/register

The parish registers differ in many ways from the censuses. A census is carried out at a specific date, it uses a pre-printed form and it is generally carried out within the same geographic boundaries over time. The local area for a census is the parish and consequently the geography for a census and a parish register should be the same. This is the situation in many cases but very often a vicar kept a register for more than one parish. To make matters even more complicated this might change when a new vicar was appointed so both place and time for a parish register for a specific parish can change.

A parish register is a physical book and the vicar kept using the same book until it was finished. This would vary among the parishes depending on the amount of 'events' taking place (and the vicar's dedication to the task – especially before 1814). So the time period covered in a parish register is far from fixed and this is a problem when you want to register which parish registers have been transcribed for which period. Finally – a pre-printed form was not put into use before 1814.

Answers had to be found on what is a 'source entry'? Time, space and event types should be defined and last but not least the index fields.

The definition of time and space for a specific source entry is of special interest for the management of the source entries. The goal is the same as for the census source entry project: to avoid that two persons are transcribing the same records. As the time periods and geography vary over time extensions to the management system will have to be made. The conclusion of the analysis was that the smallest acceptable unit of a source entry should be*: 'one type of event in one parish from one parish register. In parish registers from more than one parish where the parishes are mixed up all the entries are transcribed – no selection of records from the register.*' For the structured parish registers this will normally be equivalent to transcribing one type of event from one parish register, *e.g.* all the baptisms from all the years the chosen parish register was running. For the structured parish registers this has as a consequence that several persons may transcribe separate types of events from the same parish register.

Having defined a source entry the minimum information required from a parish register in order to make the transcription useful for a variety of purposes as mentioned above had to be decided upon. The basic guideline during the analysis was that the format for the transcription should be as simple as possible and as close to being an index as possible. As will be demonstrated this was not an easy task and the amount of wanted and required minimum information increased during the analysis phase.

The initial idea was to use one schema for all parish registers and extract the information that was common for most types of events. And there should be no distinction between 'chronological' parish registers and the structured ones. This turned out to be too simple a structure and this principle was abandoned.

Another principle that was analysed was to take more consideration to the change of the structure of the parish registers over time. This would give three time periods with app. 4 types of events for each period: baptisms, confirmations, marriages, burials and in addition: in- and out-migrations. For each type and period was made a list of possible information that could be found in the sources and from this list was decided what was optional and what was required. The goal of defining only an index versus getting as much information from the sources as possible once you were actually transcribing complicated the analysis and decisions. The number of needed fields kept growing and the number of screens was too high. This structure served the needs of the professional historians but was not taking particular consideration to the user doing the transcription.

The present solution focuses on the types of events and the information you need for each event in order to create an index with sufficient information to be immediately useful for all user groups but not requiring all information from the source to be entered. This structure does not distinguish between structured and 'chronological' parish registers. When transcribing a

'chronological' parish register you will have to select event for each record but the benefit is that every parish register ends up as structured entries.

The final principal decision was to decide upon whether or not there should be any required fields. From having no required fields – which could in reality result in having entries missing some very essential information like the name of the central person or having just one field entered for each event – it was decided that *name* and *sex* is mandatory for all events except the type 'Other events'. All other fields are optional. The amount of optional fields is quite large and has ended up giving possibility for entering nearly all the information from any type of event. As an example the fields for baptism can be used.

Information on the baptised child:
date for baptism in church: both as text field and as calendar date;
Date of birth: church date and calendar date; name ; sex; legitimate/illegitimate; twin/triplet;
Information on the father: name; occupation; place name;
Information on mother; name; occupation; place name; age;
If there are any godfathers you can enter the following information:
Name; sex; occupation; place name;

In addition to this is the general information described below. In the format definitions from 1993 175 fields were defined. In the present format 211 fields are defined plus the fields needed for documentation and administration. Though the number may seem big the expectation is that as only two fields are required it will not be a problem for the users. On the contrary the fields selected are the ones that hold the most important and central information from the parish registers and therefore it will be logical to enter this information as well.

The types of event were extended to the following: baptisms, godparents, introductions, confirmations, marriages/ betrothals, best men, burials, absolutions, other events, in-migration and out-migration.

As discussed above a transcription is never the same as the physical source but still you can have the ambition of making the transcription as true to the source as possible. For the census entry project some pragmatic decisions have been taken and the decision was to follow the practice from that project. This means that names are entered as they are written in the source. Two issues must be considered here which are specific for the parish registers. As the name is mandatory but is often not registered by the vicar who has instead written *e.g.* ' the shepherd from Haurum' this text must be written as the name. Shepherd is also entered as the occupation and Haurum as the place name. If a vicar has entered: 'Confirmation of 14 boys, 7 girls' this must be entered in the name field and the sex will be registered as unknown.

The problems with dates are not trivial. Dates can be entered as the vicar wrote them using the names of the dates according to the church and they can be converted to the 'normal' date format: DD/MM/YYYY. And if the vicar has written the date using the church calendar there will be a problem in December where the church year begins. It is not mandatory to enter both the church date and the converted date. This is decided so because there might be problems converting the date – on the other hand a normalised date is immediately understandable and searchable in any system.

Source references and management

In order to make the transcribed parish registers generally useful clear references to the original source must be added. The following information is considered necessary in order to identify a source: the unique signature for the parish register, whether the parish register is kept by the vicar or the parish clerk (after 1814), the county, district, parish(-es), period and which parts of the parish register are entered (baptisms, all, etc.).

If the vicar or somebody else has added general comments about the register this information must also be added; like pages covering a specific period that are missing in the register. Further information can be added in a separate comment field if *e.g.* the volunteer has special information about the parish.

In order to identify a record within a source more information is required. For each record a record number must be added. Clear reference must be made to the number of the microfiche and the page number in the source and the calendar year must also be added.

For each record two fields are defined where comments can be entered. One field where comments belonging to the source can be entered and another field where the volunteer can enter his/her comments. This

way it will be easy to distinguish the content and type of the comments.

As the transcribed sources are parts of a larger project each entry must be added its own unique number. The administration of these numbers will be done by the Danish Data Archive. The management system will have to be extended due to the described problems with time periods and geography.

Based on the experience from the census entry project and from the decisions on required and optional fields the ownership to the transcribed sources had to be clearly defined. One scenario could be that a person had only entered the required fields for each event type. Another person finds this to be too little and wants to add more information. To make this happen the first person must not have exclusive rights to the entry when he/she has sent the transcribed data to the DDA. Consequently one, two or three persons may be actively involved in the transcription, which must be clearly documented. If possible the person who begun the transcription is informed if another person wants to add information.

The final thing to define was the data exchange. As it was not the intention to develop a program but to define formats and principles the format for the data should be as general as possible. The most generally usable format for data exchange is csv-files so this file type was decided. In order to keep the achieved structure each event type will be stored in a separate file. For the 'chronological' parish registers this will result in a number of csv-files all having the same entry number but being distinguished by the addition of event type to the filename.

The future

The formats for transcribing the parish register have been developed by a group of experts – professional historians, genealogists and data archival experts. This should guarantee that the central information is selected in order to be useful for both researchers and genealogists. The transcribed parish registers can be used for both genealogy but certainly as well for demographic and statistical purposes.

The next step is to make a real test of the formats by having persons outside the group test the formats against the sources.

With this project a general principle and method for defining indexes for other types of sources has been developed and it is to be expected that more types of structured sources will be transcribed within the next years.

References

DDA Nyt Danish Data Archive no. 65 1993

H. Chr. Johansen: *Danish Population History 1600-1939.* Denmark 2002

Historiske massekilder, Hans Jørgen Marker, Fortid og Nutid 1993

Is there a role for data warehousing technology in historical research?

J. Delve & R. G. Healey***

Background

The harbinger of this paper was the review by Delve and Allen[1] delivered at AHC2003 in Tromso, Norway, of the potential offered by data warehousing for historical research. Essentially, data warehouses (DWs) are built for the analysis of large quantities of varied types of data, and have been used in industry extensively over the last 15 years or so[2]. Given their corporate home, it is not surprising that the attendant methodologies for DWs are geared towards business in general and sales in particular. In essence, DWs are massive repositories of clean data taken from a variety of sources – databases, spreadsheets, text files and the like, which can then be integrated or summarised as required. As the DW is so heavily data-oriented, meta data is of paramount importance and is harvested at every stage of the warehousing process. Coding of any kind is discouraged[3] as clear, primitive, well-documented data forms the bedrock of any DW, thus ensuring future usability and interoperability. The industry is currently divided about the DW lifecycle. Inmon advocates a data-driven approach (similar to Thaller's source-oriented ideology[4]), which makes no prior assumptions about analytical needs or requirements. A large DW is created and then used to populate individual data marts (DMs) for each department. These DMs are also called hypercubes, OLAP cubes or multidimensional databases (MDBMSs). They typically contain mostly aggregate data, and their main function is the analysis and visualisation of data. Conversely, Kimball recommends determining the business needs, then building up a series of DMs which must be carefully integrated to form a DW. Inmon emphasises the fact that the DW relies on many different technologies and should be thought of as an 'architecture'.

There are many differences between DWs (and DMs) and relational databases. One key distinction is that DWs are not normally updated – they just have fresh data added to them, so that the normalisation rules for relational databases do not apply. A very useful spin-off from this is that indexes are easier to maintain, and these form the DW's analytical powerhouse. The data model for DMs and DWs is the dimensional model, comprising a fact table surrounded by dimension tables to form a star schema. The fact table typically stores numerical data, whilst the dimension tables hold textual descriptors, often set out in a hierarchical manner. Kimball has recently branched out, however, creating the university DW which has a 'factless' fact table which is focussed on textual rather than numerical data. As the dimensional model has been specifically conceived of for analytical purposes, it lends itself to semantic modelling much more readily than the relational model. The storage of such large amounts of data has necessitated the development of new forms, and sparse data handlers are used for the optimisation of null-rich data.

Historical researchers are starting to harness the power offered by DMs / DWs, with the Canadian Century Research Infrastructure[5] (also discussed at AHC2003) aiming to house census data from the last 100 years in DMs constructed using IBM software at several sites based in universities across the country. At the University of Portsmouth, UK, a historical DW of American mining data has been constructed using Oracle Warehouse Builder[6]. This is examined in more detail below. These projects give some idea of the scale of project a DW can cope with, that is, really large country / state -wide problems.

* History of Computing Group, School of Computing, University of Portsmouth
** School of Computing and Department of Geography, University of Portsmouth

Models of information management

The fundamental issue in the adoption of data warehousing technology for educational/research data resource management is the model of information management required. Data warehousing is associated with a corporate model, whereas research is much more aligned with an individualistic model.

The corporate model has the advantage that a specific person, such as a chief information officer, or a small group of senior managers, can be charged with setting policies and standards for data collection. Crucially, they can instigate changes in procedures, if required, to ensure compliance with such standards and policies. This ability to direct and channel organisational activity, to maximise the value derivable from linking related data sets, is a signal advantage for the subsequent exploitation of data warehousing technology.

In contrast, the individualistic model of research data collection, suffers from a number of inherent disadvantages:

1. There is no central coordinating body, with power to identify or enforce data standards.
2. The nature of research is such that in the historical domain, the final structure and characteristics of the data may not be known until the completion of the study. At this point, there are no remaining funds to 'retrofit' compliance with standards, which might themselves fail to take account of the special nature of the datasets derived from the project.
3. Coding/classification systems may be project specific. One investigator may disagree with the approach adopted in earlier studies, and certainly may not wish to expend scarce project resources on ensuring inter-operability with other datasets not required for project deliverables.
4. Providing metadata may be regarded as a 'low-yield' activity, the perspective of the researcher may not mesh with that of the librarian/information scientist.
5. Temporal and spatial frameworks may be project specific. Data could be available by decade, annually, monthly or for irregularly spaced time intervals. Coverage of the relevant topics could be partial and overlapping, *e.g.* production data for a large group of firms on an annual basis, together with a subset of more detailed data for a partially overlapping group of firms on a monthly basis. Equivalent problems frequently arise with the spatial units for which data are available. These may vary in boundary location or degree of aggregation from one time period to the next, as described by Gregory et al.[7].
6. Data cleansing, that is appropriate for a single project may be inadequate or inappropriate when linkage to data sets from other projects is required. AHDS standards may help here, see Townsend et al.[8]
7. Gaps in data coverage from individual projects may inhibit linkage between projects, or produce an overall data 'matrix', which is very sparse.

Although the above issues may cause a number of difficulties for a co-ordinated DW approach, this should not be taken to imply that adopting a DW perspective, even if pursued informally, would be a fruitless exercise from a scholarly point of view. Assessing the degree of possible linkage between datasets is an indication of the potential 'added value' to be derived from investing time in such activity. On this basis, some linkage projects may be deemed worthwhile in academic 'cost-benefit' terms, while others would be unsuitable. A further group might indicate potential for linkage, if additional small projects could be identified, that would provide the necessary linking data tables. Even identification of gaps in overall coverage, following evaluation of multiple data sets, has a positive function, in focusing attention on areas where future priorities for research investigation might lie.

Such arguments suggest that a DW approach has the potential to foster a more collaborative approach to the collection and structuring of historical data resources than a more limited data archiving methodology. The latter focuses quite correctly, on issues such as copyright, access rights and acquisition of metadata, i.e. metadata about provenance, changes to data etc. However, content metadata tends to be more limited and more variable in quality, whereas this is the starting point for data warehousing. The latter would also necessarily be more active as a methodology, in the sense that linkages between project resources are actively sought, whereas data archiving is a more passive undertaking, dependent on the goodwill of individual investigators, or contractual obligations on grant-holders to supply data resources after the conclusion of specific projects. The ability of a DW to house disparate data sources is surely an advantage here.

While proponents of DW are keen to stress that it is an architecture rather than a technology, as already noted, there are nevertheless a number of related tech-

nologies that are quite specific to DW applications, to the point where many of the operations involved are supported by extensions to the SQL query language. The most important of these are associated with online analytical processing methods (OLAP), and include SQL operators such as ROLLUP, and CUBE[9]. Such operators would not be employed in a standard relational database application, since they require the data tables and associated metadata to be set up in the form of fact and linked dimension tables, i.e. that a DW has previously been created. Other technologies include materialised views and bitmap indexes, both designed to improve the performance of queries against DWs. Further technological extensions, which can build on the existence of a DW would include data mining and use of modelling sub-languages for scenario forecasting, although these are beyond the scope of the present paper.

The implications of these linked technologies are quite clear, however. The archival and data cataloguing approach, very valuable though it is, is essentially limited to manipulation of dataset metadata in a restricted manner. The DW approach extends far beyond this into sophisticated retrieval, manipulation, and even analysis of the contents of the data sources themselves. In leading proprietary toolkits, such as Oracle Warehouse Builder, not only are these facilities available to the user, but they are also callable from a Java application programming interface. Such programs, via other libraries, can also link directly to the Web, providing a direct live connection between the global research community and the DW. Hence the future potential of this approach for facilitating and enabling scholarly endeavour is far greater than that of current archival methods.

Data warehouses and multi-media content

A further issue that is of growing importance is the ability of DWs to handle data other than the standard character, numeric and date data types for which they were originally designed. Such data would include free text, scanned images, sound, video, and indeed digital map data, describing points, lines and polygon features. However, this question needs to be approached with some caution since, in the case of multimedia data, at least, it is not immediately obvious that OLAP methods are relevant to the internal contents of digital video or sound clips, apart from contrived queries, such as give me a breakdown by film studio of all video clips featuring actors, whose body shape resembles that of John Wayne! It is much more likely that the query would be resolved in the descriptive fields linked to the clip library. The results set could then be examined by the enquirer at their own expense, doubtless on a pay-per-view basis!

The situation is slightly different for free text data, which might take the form of electronic books, newspaper articles or transcribed manuscript materials. While the bibliographic metadata are certainly important and would probably be the focus of a first stage query, we are likely to be just as interested in the location of specific words or phrases in the body of the texts and the extent of our word search may be refined or amplified using thesaurus type facilities. Certain types of users could be interested in queries concerned with word frequencies in texts purporting to have been written by specific authors between nominated dates. Others may be trying to identify groups of documents with the highest yield, in terms of the greatest frequency of occurrence of key phrases. Text retrieval is not, of itself, a standard DW or OLAP operation, although small text fragments could be stored in normal character columns in a fact table, if required, and text indexed there. It is interesting to note that dimensions can encapsulate overlapping hierarchical structures, in a similar fashion to broader and narrower terms in a thesaurus. It would be possible, therefore, to set up a thesaurus in dimensional form, although the star schema indexing requires numeric keys, so the word tokens in the dimension would have to be referenced by ID numbers that linked to the record numbers in the fact table, where they could be found. This would involve significant processing work by an experienced database programmer and would only be feasible in software systems that gave users access to the text indexes themselves.

Spatial data warehouses?

The hierarchical nature of dimensional structures already allows certain kinds of geographical aggregation and disaggregation to be performed within the DW. For example, census enumeration districts could be nested within wards within boroughs within counties, and overlapping aggregations involving civil parishes for health authority areas could also be handled. Such aggregations can also be date dependent as Southall et al.[10] have shown very clearly. Importantly, OLAP reports can be generated from these hierarchies, with-

out any reference to the cartographic elements or area boundaries themselves. If reports involved GIS type queries, such as finding all points within 15 miles of the boundary of a polygon, or polygons that are adjacent to other polygons with specified characteristics, then a different approach is required. If the database system in which the warehouse resides has spatial functionality, then spatial indexes and operators will need to be involved to resolve the query in question. If not, the database will have to be interfaced to a proprietary GIS, where the necessary analysis can be undertaken.

Performing spatial queries is one issue. A second in this context is the display or visualisation of the results of queries. Some systems, such as Oracle, have basic map display front ends for showing map output, based on retrieval of both digital cartographic and attribute data held in the database. Where such facilities are not available, again recourse must be to a proprietary GIS or graphics package for output purposes. That said, the multi-dimensional nature of OLAP queries, which report summary data for different 'slices' of a multi-dimensional cube in sequence, seems to offer the potential for new types of map visualisation or animation that have yet to be explored.

Technical Issues for Incorporation of Multi-Media and Spatial Data in Warehouses

There remain three key technical issues under this heading. The first is that these data sources are now usually held in specialised unformatted 'container' datatypes, usually referred to as BLOBs or CLOBs (binary or character large objects). These datatypes are not supported by the current generation of DWs, although this may change in future. However, it is probably not worthwhile to 'shoe horn' free text or digital map coordinate data into standard numerical character fields, especially when applications become large. In the case of multimedia data this is simply not possible.

Secondly, therefore, if resources of these kinds are to be included within the warehousing framework, at present levels of technology, it seems to be essential that the warehouse sits within a relational or other suitable type of database framework, where these different data types can be stored and manipulated. Suitable metadata can then be handled within the warehouse structure and relational or pointer links made from the relevant fact table or dimension table rows to the tables outside the warehouse, but within the encompassing database, where the container data types are held. In the case of spatial data, such metadata could include bounding rectangle 'foot prints' of map polygons, as in the Alexandria Digital library[11]. From a narrow technical perspective, the integration of the warehouse with the wider database framework should mean that a query or series of programmatically linked queries can access both OLAP operators and other operators specific to the different special-purpose data types. In this context, it should be noted that 'factless' fact tables are not a solution to this issue, as they simply transfer the measure information from the fact table to the surrounding linked dimensions, rather than to tables external to the warehouse structure.

The third major technical issue is that such systems require sequential, if not simultaneous deployment of multiple indexing strategies. These include standard binary tree indexes for non-key columns in fact or dimensional tables, bitmap indexes for accelerating OLAP queries on the star schema, text indexes for free text retrieval, image bitmap indexes for image pattern recognition (if required) and multi-dimensional spatial indexes for resolving map based queries. Optimising queries on large data volumes using some, or all of these indexing methods is an extremely demanding task. Nevertheless, being able to discuss the possibilities involved is itself an indicator of the advanced technological level that the best available universal server databases and data warehouses have now reached. It is equally clear that few, if any, current repositories remotely approach the levels of sophistication that are now potentially achievable, either in terms of the range of search strategies that they can deploy, or in the extent to which they have linked different data sources with a view to maximising the 'intellectual added value' that can be derived from querying them in combination.

Lessons from a mining history case study

In the course of exploring a number of the above issues in both the research and teaching context, an operational demonstrator warehouse has been developed using Oracle Warehouse Builder technology. The application is to the anthracite coal mining industry of Pennsylvania in the 19th century, and it is based primarily on raw data previously held in a normalised relational database format, supplemented by legacy data from spreadsheets. These datasets were originally developed over many years, in the course of writing an interpretative study of the anthracite coal industry during this period.[12] The main fact table holds coal

shipment data and is linked to dimensions for specific mines, different operating companies and time, measured in years or decades. Several major lessons have already been learned in moving from a standard relational structure to a warehousing environment. The first is the importance of rigorous data cleansing. For example, the original data sources record nearly 500 variants of company names. When standardised, this number reduced by half, a result that required significant work to achieve. However, were it not done, the value of resulting queries by company would have been extremely limited. Secondly, a consistent approach to missing data is essential to allow queries to work as expected. Thirdly, OLAP queries are very powerful means of summarising segments of the data 'space'. For example, an analysis of the relationship between production levels and standardised accident rates was run, broken down by some areas of the coalfield, and this was immediately suggestive of further avenues for investigation. Fourthly, the speed and power of such queries across large data cubes, focuses the mind on the requirement for reliable and accurate data, otherwise the old GIGO principle could be elevated to entirely new heights!

At present, the warehouse is still project specific. However, inclusion of other candidate dimensions, such as shipping railroads, would allow the possibility of linkage to other projects, *e.g.* on the iron industry. The incorporation of spatial data for mine or railroad locations, which is already available, but not yet in a database format, would also allow the question of map based results from warehouse queries to be explored.

Conclusions

Warehousing as a data architecture has much to commend it. At the broadest level, it imposes a valuable discipline on all stages of data preparation, linkage, retrieval and analysis. At a more detailed level, the use of the fact and dimension table approach enables OLAP tools, some of the most powerful weapons in the analyst's armoury, to be deployed. In terms of specific technology, the leading vendor's provide graphics based design and management tools for warehousing applications that allow many stages of the process to be semi-automated, thereby improving productivity. However, there are compelling arguments, in respect of linking a warehouse structure to multimedia content, for using warehouse systems embedded in universal server type databases. For small projects, warehousing is overkill, because of the personnel resource overheads in establishing the necessary infrastructure. For large projects, or medium-sized projects with complex data structure, the return on investment should be positive.

The final question of the corporate versus the individualistic model of archive or data resource management remains as a sticking point. In the academic world, we have already reached the point where the value of our digital data vastly exceeds the capital cost of the hardware and software used for data storage. However, a future champion of the warehousing cause will still need to convince a wide variety of stakeholders, that a move to the more directed acquisition and linkage of data sets is justified, in the face of competing calls on the scarce resources of research time and funding.

Notes

1 An article is in progress for *History and Computing*, but in the meanwhile, see a related paper aimed at the IS community – 'Data Warehousing – the new Knowledge Management Architecture for Humanities Research?', May 2004, UKAIS (United Kingdom Academy for Information Systems) Conference Proceedings.

2 Inmon, W. H. (2002). Building the Data Warehouse, New York: Wiley

3 Kimball, R. and Ross, M. (2002). The Data Warehouse Toolkit. New York: Wiley, p21.

4 Thaller, M. 1991 'The Historical Workstation Project', *Computers and the Humanities,* **25**, 149-62, p155.

5 See www.canada.uottawa.ca/ccri/

6 Delve, J., Healey, R., Fletcher, A. (2004). 'Teaching Data Warehousing as a Standalone Unit using Oracle Warehouse Builder', *Proceedings of the British Computer Society TLAD Conference,* July 2004, Edinburgh, UK.

7 Gregory, I.N., Dorling, D. and Southall, H.R. (2001) 'A Century of Inequality in England and Wales using Standardised Geographical Units', *Area*, 33,297-311.

8 Townsend, S., Chappell, C. and Struijve, O. (1999) Digitising History, A guide to Creating Digital Resources from Historical Documents. Http://hds.essex.ac.uk/g2gp/digitising_history/index.asp.

9 ORACLE Corporation (2003) ORACLE Database SQL Reference 10g Release 1. Redwood City, CA: ORACLE Corporation.

10 See the census data retrieval facilities in *www.visionofbritain.org.uk*.
11 See www.alexandria.ucsb.edu.
12 Healey, R. G. (forthcoming) From the Civil War to the 1902 Coal Strike: Recession and Resurgence in the Pennsylvanian Anthracite Coal Industry. Scranton, PA: University of Scranton Press.

Computer science and the Dutch cultural heritage

*Paul M. Doorenbosch**

1 Introduction[1]

The collective memory of the Netherlands is stored in our cultural heritage. Huge numbers of archives, books and magazines, paintings and other art objects, audio-visual resources, folkloristic and archaeological objects are kept in numerous different places, often in buildings that are themselves part of our cultural heritage. These objects are witnesses of our past and present. They are indispensable components of our national identity and our society. They contribute to the quality of life. Nowadays, a growing number of people – teachers, researchers, politicians or people with a general interest in culture – prefer to have access to these cultural collections, independent of time or location: at home or in the street, on their mobile phone or on their computer. It would also be very convenient if these collections could be approached as one integrated collection: the Dutch cultural collection. To accomplish this, museums, archives and libraries try to realise effective access to digitised objects and digital durability of our entire cultural heritage.

Information Technology (IT) is the ideal companion in this venture. Although the level of IT is different in every sector of cultural heritage, it is widely agreed upon that IT is an indispensable part of managing and presenting cultural heritage nowadays. Most of the institutes are too small to run a research and development department to follow and implement actual developments in the field of IT. It is nearly impossible to attract funds for internet development. Measurement of success is still based on the numbers of actual visitors to the location. Sponsors are interested in a location they can invite their relations to, not in an application on the internet or a project to enrich existing data.

2 CATCH (Continuous Access to Cultural Heritage)

NWO, the Netherlands Organisation for Scientific Research (www.nwo.nl)has started a programme in which computer science and cultural heritage work together to increase the semantic access to digitised cultural heritage in the Netherlands and to develop scientifically relevant methods to acquire new fundamental and applied knowledge about this process. This programme, called CATCH (Continuous Access to Cultural Heritage) (www.nwo.nl/catch), started in 2005 with six core projects. Researchers from seven IT knowledge institutes – mostly university research groups [3] – and six leading cultural institutions [4] are co-operating in executing projects within the three research themes of CATCH:

– semantic interoperability through metadata
– knowledge enrichment through automated analyses
– personalisation through presentation

An open call in which new combinations of IT and cultural heritage institutions can submit proposals is now under construction and will results in four new projects. It is expected that another open call will follow. Funding is for the greater part provided by NWO.

The four main objectives of the CATCH programme are:

– To discover new methods, tools and computer science knowledge in different IT-fields, such as Semantic Web, agent technology, artificial intelligence, and knowledge enrichment.
– To increase the efficacy and efficiency of digital access to Dutch cultural heritage for research, education, and the general public.
– To reinforce the knowledge infrastructure of both IT researchers and the cultural heritage sector

*Koninklijke Bibliotheek, National Library of the Netherlands, The Hague

– To improve the international position of Dutch researchers within the interdisciplinary research field of computer science and humanities.

The volume of the Dutch digitised cultural heritage is huge and is increasing every day. The use of advanced information and communication technology enables sustainable accessibility of this digitised cultural heritage. The purpose of the CATCH programme is to carry out fundamental and applied research. Besides the development of new knowledge and methods, it will result in a framework of tools, that should become part of the overall knowledge infrastructure providing the public with direct, time and location independent, access to the greater part of the digitised Dutch cultural heritage, and answers that are relevant and meaningful in the context of their queries.

The potential users of the results of CATCH fall into two categories.

– The collection managers of the cultural heritage institutes.
– The end users of the services provided by the cultural heritage institutes.

Many of the researchers involved participated in earlier joint IT and cultural heritage projects, such as ToKeN 2000 (www.token2000.nl). The huge amount of data as well as its diversity makes cultural heritage a challenging IT research domain. The main difference between CATCH and earlier research programmes is that CATCH is specifically based on questions from the heritage sector (*demand pull)* while the other programmes are mainly technology-driven.

2.1 The cultural heritage sector

In the cultural heritage sector we distinguish five main sections, namely museums, libraries, archives, monument services and archaeological institutions. Obviously, the aims and intentions of these sections are closely connected and rather concurrent, especially in the way they try to attract and 'entertain' visitors. However, with respect to IT developments, and in particular in the effective and efficient application of IT tools, their goals, organisational methods, and stages of development diverge considerably.

During the last decade cultural heritage-institutions digitised (parts of) their collections. At first these digitised collections were relatively small and simple in structure. This resulted in a motley collection of digitised objects, with different kinds of gateways to the Internet and with a variety of access paths from the perspective of the user. Moreover, the distributed character of the digitising process has led to scattered storage and varying accessibility procedures. The past few years have seen increasing investments in cross-sectoral projects, such as the Memory of the Netherlands (www.geheugenvannederland.nl), carried out by the national library of the Netherlands, Koninklijke Bibliotheek (www.kb.nl), which is basically an imaging programme (from analogue to digital), and the Cultuurwijzer/Cultuurwijs project by the Netherlands Digital Heritage Association, DEN (www.den.nl) [2]. 'Cultuurwijzer' (www.cultuurwijzer.nl) is a selective portal to a large number of digital heritage information resources for the general public. It provides a generic structure and solutions for improved access to digitised cultural heritage collections, exceeding the limits of the individual collections. The latest addition is KICH (Knowledge Infrastructure Cultural History), which aims at connecting information about cultural history from different sources, mainly geographically based (www.kich.nl). The integration of knowledge stored in different databases is still in its infancy in all these projects. Mostly they are just portals, or they provide integrated search facilities with hyperlinking. But the tendency towards real knowledge integration and knowledge enhancement is visible.

2.2 Interdisciplinarity

One of the focus points of NWO's strategy for the coming years is interdisciplinarity. Interdisciplinarity is inherent in the overall design of the CATCH programme, since it is defined as an IT programme for the cultural heritage sector. Interdisciplinarity plays a part in all the subjects addressed, in the various backgrounds of the researchers involved, and in the way the programme and its individual projects are implemented. It is considered to be an essential feature of the programme that a large part of each of the individual projects is carried out at the locations where the cultural objects are actually stored (or exhibited), i.e. in the museums, libraries and archives. This implies that the researchers involved will be based in a cultural heritage environment and will be supervised and advised by a cultural heritage expert on a daily basis.

The projects within each research line are focused primarily on that specific research line but will also be strongly connected to the projects in the other research lines. An intensive schedule of meetings and a protocol of how to exchange information will strengthen the co-operation.

2.3 Metadata

With respect to topical metadata one inherent problem is clear: the meaning of terminology changes. Hence, in the future, concepts with the same meaning may be referred to by different terms. Moreover, the use of a classification system will change in the course of time, due to its growth and the development of scientific insights. For a long time cultural heritage researchers believed that clarifying the connection between older records and newer ones and indicating the association between historically connected objects could only be performed manually (at least for the greater part). The CATCH programme will show that it is possible to develop tools to support the user in making historical connections at the topical level.

In the case of retro-digitisation there is also a lack of interconnectedness at the level of the descriptive metadata. Standards for the description of objects in older catalogues and inventory systems deviate from the current standards or are completely lacking, resulting in poor connections. Nowadays, when objects are digitised, their descriptions must be carefully documented. It is of vital importance that the researchers know the standards according to which the catalogues were compiled, as well as the changes made over the course of time. The need for a relatively new kind of metadata is also felt: administrative, technical and preservation metadata, describing the procedural and technical aspects of the digitisation process itself. A number of important developments in the field of metadata and the preservation of digital objects have already taken place in the library and archival world.

Some of the CATCH projects will be geared towards examining the vast knowledge bases of the digital heritage in order to try to automatically identify and categorize the various knowledge elements. Due to the heterogeneity of formats and conventions, indexing and classifying relevant knowledge is a difficult task, technically as well as semantically, but, according to the project leaders of CATCH, not an impossible one.

Of all the research lines, metadata is most closely related to the established practice of curators and other cultural heritage workers.

2.45 Unlimited knowledge enhancement

The main goal of cultural heritage research is to discover new knowledge, to add to the existing body of knowledge about the objects and their creation processes. Knowledge enhancement is a relatively new research topic that originally focused on generating new knowledge from existing knowledge. Nowadays, knowledge enhancement is supported by a plethora of techniques, such as neural networks, pattern classifiers, data mining, etc.

The object of knowledge enhancement is to clarify how objects, information, or links are to be interpreted. It may deal with a certain type of object (for instance a certain painting which belongs to the impressionist school of 1880-1890), but also with types of material, contemporaries, and other, not immediately obvious relations. Thus, new relations encompassing diverse objects from different collections may become apparent. Automatic knowledge processing techniques may even take over parts of the research task, in particular when very large quantities of object data are involved. Knowledge processing techniques will find new relations in unstructured object data collections and object databases. To do so it is essential to have techniques that are able (1) to scale up to very large databases, (2) to produce overviews, analyses, and indices that can be inspected and understood on a human scale, and (3) to normalise, enrich, and clean up very large databases that could not be normalised by humans on a human time-scale. The CATCH research line will focus on state of the art work in machine learning, a subfield of artificial intelligence. Considerable scientific innovation is furthermore expected to arise from investigating the very large scale of cultural heritage data in combination with the rich multi-modality (sound, image, text) and relational structures (thesauri, history) in these data.

2.5 Navigation and presentation: ubiquitous accessibility

A principal aim of CATCH is exploring methods and techniques to increase the efficacy and efficiency of digital access to the Dutch cultural heritage. Achieving this aim is considered to be a major advance, for which many smaller scientific and technological improvements have to be accomplished. As a case in point, we take the metadata research line. To get access to an art object, an agent acting on behalf of the user should be able to navigate to the correct location. Metadata serve as signposts in the navigation process. The better the metadata, the better the results in terms of precision, recall, speed, and reliability. Moreover, metadata can pave the way with respect to interoperability obstacles. Furthermore, metadata can support

presentation, semantic-based access, and knowledge enhancement. Hence, improvements in the definition and use of metadata are needed for breakthroughs towards ubiquitous accessibility.

When navigating the cultural heritage knowledge bases, users are invited to translate their personal needs and desires into digital actions. Presenting the – intermediate or final – results in a personalised way is therefore considered to be an essential component. Once the previous stages have been completed successfully, in other words, when metadata has been added to the cultural heritage objects, when the databases are interoperable and the knowledge contained in them can be accessed in a semantically meaningful way, efficient navigation and presentation techniques provide the final stage of making the cultural heritage accessible to various groups in society (researchers, experts and the general public).

The answers to the questions posed by a user should consist of tailored information. It is therefore necessary for a computer to know what the needs of the user are. Hence, information on the user's interests needs to be gathered. The existing methods to do so ultimately fall into one of two categories: (1) those that feature obtrusive interaction with the user, and (2) those that feature unobtrusive observation. The main drawback regarding obtrusive methods is that they burden the user with additional actions. In regard to the unobtrusive observation the design of the user interface is a major issue. In order to gather knowledge effectively about the user three essential questions need to be answered: (1) which user behaviour should be observed? (2) where and when should the user be observed? and (3) how can the observed behaviour be interpreted?

2.6 Architecture

CATCH will provide a dynamic, distributed, service-based environment, comprising autonomous cultural heritage entities and service providers which make enhanced content available. The cultural heritage institutions will utilise content enhancement services to obtain additional descriptive metadata to improve the accessibility of their content. The use of open standards will ensure that the results in CATCH can be used by the entire heritage sector, that the results can be enhanced and extended by additional projects and that the programme can easily be extended with new partners. Effort put into an exploitation plan for the longer term will support the continuous usability of the results.

2.7 The six core projects

The CATCH programme will initially start off with six projects. A brief overview:

2.7.1 STITCH: SemanTic Interoperability To access Cultural Heritage

The prime research objective of this subproject is to develop theory, methods and tools that will allow metadata interoperability through semantic links between the vocabularies. This research challenge is similar to what is called the 'ontology mapping' problem in ontology research.

The overall objective can be divided into three research questions:

– What kind of semantic links can be identified?
– Which methods and tools can support manual and semi-automatic identification of semantic links between vocabularies?
– How can such semantic links be employed to enable interoperable access to multiple collections indexed with heterogeneous vocabularies?

2.7.2 CHOICE: CHarting the informatiOn landscape employIng ContExt information

CHOICE is focusing on semi-automatic semantic annotation and on employing context information. Semantic annotation involves the annotation of archived objects with semantic categories from standardised metadata repositories, such as domain thesauri and ontologies. The use of semantic annotation allows for a broadening of the search facilities in a collection. The following research issues need to be tackled:

– Which thesauri and/or ontologies can be used as repositories of relevant semantic categories for archive search and how can these thesauri/ontologies be partially mapped/integrated?
– How can we use NLP and learning techniques to derive relevant semantic categories from the text and how can these semantic categorization techniques be used to support the search process?

2.7.3 RICH: Reading Images in the Cultural Heritage

The amount of recovered archaeological objects is so vast that it is beyond our imagination. From this collection a corpus of knowledge has been built on the distribution of these objects in space and time, the evolution of ancient technology, and the function and role of particular objects in ancient society. To communicate this information, archaeologists traditionally

use the concept of reference collections: they classify their finds in types and series of types. This is a mental process which combines and recombines evidence and theory from the finds at hand with earlier archaeological research.

The field of digital vision has developed to such extent, that it now becomes realistic to incorporate these new techniques into the electronic archaeological reference collection. This will enhance the quality of archaeological research and archaeological heritage management in a fundamental way. The RICH-project focuses on automatic recognition of form, fabric, and decoration of physical objects and of printed images.

2.7.4 SCRATCH: Script Analysis Tools for Cultural Heritage

Large collections of handwritten material do not easily lend themselves for simple access on the basis of keywords or traditional information-retrieval methods. It is difficult to read the handwriting of another person, and this is even more difficult if the writing originates from a different period in history. Under these conditions it may be appreciated that the automatic recognition of handwriting, i.e. the automatic conversion of a text image to a coded representation, is an important research area. If the goals are defined realistically, current technology can play an important role in providing tools for retrieval and semi-automated methods of script annotation.

The research question SCRATCH means to address is whether the application of search and semi-automated annotation tools will provide an effective improvement of purely manual transcription methods. Furthermore, current methods in handwriting analysis systems may allow for new ways of search and retrieval. Traditionally, there is a focus on the textual content of a manuscript, but advanced techniques may allow for a detailed in-depth analysis of the character shapes as well. In order to solve the problems of accessing handwritten document collections, methods from several research domains need to be incorporated: image processing, pattern recognition for shape classification, layout modelling, content metadata research, and stochastic modelling methods from computer linguistics.

2.7.5 MITCH: Mining for Information in Texts from Cultural Heritage.

Text mining, a research domain of natural language engineering, has advanced to a level at which automatic language technology and information extraction modules can be applied to vast amounts of text. The modules analyse these texts on syntax, document structure, and topical-semantic information. There is nothing intrinsic bound to the type of text that could be analysed by these methods. The text can be 'ungrammatical' or can even be a list of terms stored in database records. The more structure a collection of texts has, the more possibilities there are for machine learning systems to learn the regularities or syntax of the structure, and apply it to new data or find inconsistencies in existing structured data.

There are also no intrinsic restrictions on the morphological, syntactic, and semantic structures that can be learned: the structures can be of a general nature, or be tailored to a domain in which particular types of entities and facts should be found and labelled. Once these generic or specific methods are developed, they can be used for supporting further annotation, fully automatic annotation, or the automatic discovery of inconsistencies in previously labelled material.

The research question of this project thus reads: how can language technology and text technology support the automation of knowledge enrichment and the understanding of digitised cultural-heritage texts and textual object databases? How scalable and robust are these techniques in analysing sentence and text structure syntactically and semantically?

2.7.6 CHIP: Cultural Heritage Information Personalization

The cultural heritage collections and relevant contextual information are distributed over many different institutes. CATCH aims to make the barriers between these institutes disappear by providing virtual integration of these collections. The CHIP project focuses on the interaction of the users with the combined cultural heritage content, and in particular on personalized presentation and navigation. Cultural collections do not consist of a discrete database of art objects. They come with a story that connects different art objects, details of objects, and background information together. Together these stories impose a structure on the collections. The CHIP project aims to present art objects by organizing them in groups, and more generally showing the relations between art objects, based on the objects' metadata, and indirectly on their links with background information.

Closely related to imposing structure is the problem of navigating through the heterogeneous collections of the institutions. CHIP does not want to restrict the user to following a predefined path; however, a navigation structure is needed to prevent the user from getting lost. A combination of searching and browsing is envisioned. Especially visually oriented browsing will be considered as an integral part of the presentation.

Personalization takes place in two stages: initially through the definition of stereotype user groups, and later through adaptation to individual user characteristics.

Personalization requires user profiles to be constructed. Cultural heritage institutions are interested in being able to reuse these profiles to understand their visitors better, both on the web and when physically visiting the museum or institute.

3 MultimediaN

CATCH is not the only computer science research programme in which cultural heritage plays an important role. 'Multimedia Netherlands' (www.multimedian.nl) deals with different kinds of multimedia and one of its application areas is cultural heritage. This programme is much more technology-driven compared to CATCH. That is why its founding fathers expect knowledge acquired in this programme will be applicable in different sectors, amongst which culture.

The objective of the cultural heritage-application project is the development of a set of e-culture demonstrators providing multimedia access to distributed collections of cultural heritage objects. The demonstrators intend to show various levels of syntactic and semantic interoperability between collections and various types of personalized and context-dependent presentation generation.

The second MultimediaN-project in which cultural heritage plays an important role is the semantic multimedia access project. It concentrates on the development of generic technology that satisfies multimedia information on a semantic level. Any search system's comprehension of a user's information need is necessarily incomplete, as this would require understanding completely the user's goals as well as the user's perception of retrieved objects. The researchers in this project investigate how, in spite of this uncertainty, effective search strategies can be offered, exploiting the following search parameters: the collection (domain knowledge, background knowledge, language, format, etc.), the user (including use scenario, interaction type, history, preference) and the system (security aspects, performance). Users often know things about the multimedia objects in a collection being searched; yet, it remains a challenge how to exploit and adapt to such background knowledge during search. In this project a search engine generator will be developed, based on probabilistic retrieval models. The project will focus on technology and tools applicable in the domain of media and e-culture and ambient settings.

Despite the different points of view (demand-driven versus technology-driven) CATCH and MultimediaN do explore the same field. Both parties have therefore expressed their intention to exchange knowledge and results.

4 Challenging questions in cultural heritage for computer science

CATCH and MultimediaN have a scope of four to six years, at the end of which it is highly unlikely that all cultural heritage problems will be resolved, even if the results of these programmes are re-engineered into actual applications. Although commercial companies and vendors will play a role in the development of these applications, CATCH means to find ways to keep the core software available as open source for the cultural heritage communion.

There will remain many challenging questions in which computer science and cultural heritage can reinforce each other. Let me conclude by exploring some of these topics.

– Metadata will become increasingly essential to access and preservation of digital heritage objects and knowledge, while the intellectual effort to create it takes up a large part of the available time and money. It is likely that this will become less acceptable in the future from a financial and managerial point of view. Cultural heritage will continually need new tools and methods either to create metadata (semi-) automatically, or to retrieve objects and knowledge without creating metadata beforehand.
– Text plays an essential role in the preservation and dissemination of knowledge. The contemporary means for text search and retrieval are still fairly simple. The language technology sector conducts research into the analysis of texts in order to navigate through and between texts, and detect and mine knowledge stored in these texts. This research area is very promising. The Dutch-Flemish STEVIN

programme (www.taalunieversum.org/stevin) is a research programme worth mentioning in this context.
- What will the significance of the Semantic Web be to cultural heritage? The essence in history and culture is its semantics, the meaning of it in different contexts, in different domains. It is currently impossible to handle this meaning satisfactorily in a digital environment. The Semantic Web may bring new tools and techniques to cope with the meaning of knowledge. The Semantic Web looks promising and is definitely worth investing in.
- Digital durability, long term accessibility, may well be the biggest challenge faced in the coming decade. If all knowledge is stored in a digital environment; it is essential to find ways to guarantee long term accessibility. The awareness to this problem is growing, but the solutions are still in their infancy. The question of digital durability exceeds the cultural sector, since it is relevant to all fields of society. Consequently, it is an extremely interesting field for co-operation between all sciences.
- Lastly, authority and quality no longer are what they were before. The Internet environment sometimes makes it seem as if the Internet as a whole is the new authority. How can a distinction be made between quality and non-quality, how can 'Wahrheid und Dichtung' be recognised, is it possible to easily find good or relevant information? At the moment an internet search is time consuming and does not necessarily provide relevant information.

It is a promising development that computer science and cultural heritage have found each other. Although the co-operation between these two is just a first step, it allows cultural heritage to take up a new and advanced place in contemporary society.

Notes

1 My thanks to Jaap van den Herik (IKAT, Maastricht University), Alice Dijkstra (NWO), and the other members of the central CATCH programme committee; to the scientific leaders of the CATCH core projects: Antal van den Bosch, Erik Postma, Frank van Harmelen, Lambert Schomaker, Mettina Veenstra, Paul de Braa, and the members of their teams; and to Pauline Bodde (KB).

2 At present maintained by the Netherlands Institute for Cultural Heritage (Amsterdam) (www.icn.nl)

3 The principal participating IT-research institutions are: Maastricht University, Max-Planck-Institut für Psycholinguistik (Nijmegen), Technische Universiteit Eindhoven, Telematica Instituut (Enschede), Tilburg University, University of Groningen, Vrije Universiteit (Amsterdam)

4 The principal participating cultural heritage institutions are: Dutch National Service for Archaeological Heritage (Amersfoort), Koninklijke Bibliotheek (Den Haag), Nationaal Archief (Den Haag), Naturalis (Leiden), Netherlands Institute for Sound and Vision (Hilversum), Rijksmuseum (Amsterdam)

Large longitudinal, nominative databases in historical research

*Stefan Fogelvik**

Providing the research community with historical demographic data – strategies adopted by Stockholm City Archives

Introduction

Stockholms City Archives have more than two and a half decades of experiences in digitizing large historical archives and disseminate them to different users. During these decades we have migrated our digital archives between a number of technical hardware platforms as well as made use of a number of different software systems.

We have made it our business to take advantage of these transitions to create new and better ways to disseminate and make our digital holdings useful for different users.

The creation of large historical databases often is a long term undertaking. This is something that must be considered when you decide on what strategies to adopt for the digitisation process and the way you structure the digitized information. As an archives institution we also have obligations for the long-term preservation and provision of our digital archives. That is a much bigger a challenge and undertaking than serve the users with the information structured to suite the needs of a special user-group.

When you build databases based on large longitudinal demographic archives that take decades to finalize you must provide means to make it usable during the long digitization period. This is something that's best done by working together with the users and that has been an essential guideline for our work at the Stockholm City Archives.

I will discuss the strategies that we work after at present based on our long experiences to provide our users with useful longitudinal demographic data. References will mainly be made to the digitized parts of the Roteman archives. The contents and structure of the Roteman Archives and its potential value for research has been presented at earlier AHC-conferences and in at a number of papers (1, 2, 3 and 4).

The discussion will include the use of different media and various ways of pre-processing the digitized sources to fit different kind of uses and users. We also try different forms of collaborations with the users in order to get valuable feedback for how to go ahead in our dissemination undertakings. This has been especially important for the development of user interfaces, a very important issue when we look at new ways to search, retrieve, process, analyse and present the information in our digital databases. Our interface solutions have also been very useful as part of ALM-projects where the cultural heritage is used as a source for research and education where we at present are engaged in a large project with the schools in Stockholm.

Background

Among the tasks and obligations for the Stockholm City Archives is to provide and develop routines to enable user to get access and make use of our holdings. Technological development provides us with new means and tools for these tasks at an accelerating rate. But this doesn't solve the large undertakings we have in order to digitize important collections among our large archives holdings. With limited resources we have make priorities and to guide us in doing that we need to get the opinion of our users. The Roteman Ar-

*Senior manager, Digital development, Stockholm City Archives

chives is an early example of that approach. It was the academic research community that saw the potentials of its information for research in demographic and social urban history in a spatial context. This led initially to a special project within the City Archives that is an integrated part of the Archives organization and part of the regular activities.

Longitudinal records

Today as a citizen you leave a lot of electronic traces of your activities and behaviour by making phone-calls, sending e-mails, paying bills at restaurants, stores, hospitals or theatres etc. with a card or caught by a surveying camera for speeding. Compiling information from these different digital sources it is possible to follow in your footpaths in time and space and learn a lot about your preferences and way of life. These different electronic marks you leave are already in used for analysis applicable to commercial uses. The commercial companies acknowledge the importance of a good knowledge of their customers. They are examples of longitudinal records.

A society that wants to provide its citizens with good public services needs to have a good knowledge of the socio-demographic composition of its population. The Roteman system was created in 1878 to meet the needs of the City Officials with information to be able to plan and provide housing, schools, poor relief, medical care, streets, water supply, transportation etc for the rapidly increasing population in the capital city of Sweden. The Roteman system was longitudinal and each single individual were followed during their stay in Stockholm. Apart from basic demographic information notes on profession, education, social benefits, family- and household standing, intra-city migration etc were entered as they occurred. Digitally compiled a good source to study changes in socio-economic conditions in late 19th and early 20th century Stockholm. The system ended in 1926 when changes in the city administration provided new ways to gather information on the population.

Digitization strategies

From the start we adopted a geographic approach for the digitalization work. Taking one Rote (ward) at a time for the whole time period. We also started out by using the ward as a base for storage and provision of the digital records. Since we anticipated the long time it would take to complete the digitizing we did use a system independent solution, which we so far have succeeded to migrate through a number of different technical platforms with the information intact. Today the basic Roteman files contain more than 3.6 million entries on close to one million individuals.

Our basic concern as an archives institution is to reconstruct the information as it stands in the sources in this case as alpha/numeric text. By using the structure of the Roteman archives we have and are able to reconstruct and restructure the information to meet the needs of different users in a varied way. But being true the source doesn't simplify the use of the data. The need for standardisation is evident. Since Swedish is not a natural choice as a foreign language we see international cooperation in this field is a very good way to go ahead. We have been adopting the HISCO and HISCLASS schemes on the profession and titles information in our files and also added the English versions to simplify for external users to make use and sense of our data. (www.ssa.stockholm.se).

User history

It was the academic community that took the initiative to digitize the Roteman Archives at a time when social history was very much in focus. What the research community had not anticipated was how long the undertaking would take before any substantial material could be delivered. During the wait other themes for research became more in fashion. We had anyhow a contiguous number of undergraduate students who used what digital material we could provide for their research papers. We presented what we where doing to different potential user groups. Teachers in history in secondary schools with an academic past and knowledge of the archives were the first to respond and show a genuine interest to use the digitized Archives as a source in their education. This started a long and continuing fruitful dialogue with one of our important user categories: secondary school students and tomorrow's researcher.

When we started our initial cooperation with university and secondary school students in the early 1980th there were no PC's around and we had to do most of the information processing: sample/extract, process and report/present. This meant that we had to have close contacts and discussions with our user on what they wanted and how they wanted it; a time consuming but very rewarding experience, which has been very beneficial and fruitful for the development of our serv-

ices to our users. There was however a legal obstacle for the provision of our services, the Roteman archives contains information on people still alive. We had a special law that stipulated the roles on how to handle and submitting the information outside the Archives. They have today been replaced by new regulations PUL, which stipulates the ways in which we can publish our digital files on different media and platforms.

In the late 1980s our digital holdings were of a considerable size and a new group of users started to show an interest: the genealogists. They were used to microfilm and fiche readers, so for us COM-fiche was the natural solution by which we could provide them with data and also create entry index to the Roteman information that suited their needs. We also made headings in English for the Mormons in Salt Lake City for their genealogic department.

Many of our users were also utilizing the information and services provided by other institutions in the ALM-sector (Archives, Libraries and Museums). In discussions with our colleagues in the cultural heritage realm we began talking about finding ways for cooperation in order to disseminate our holdings to our users. This was also something that got strong support from our users. After some years of informal cooperation and collaborative efforts we finally found an opportunity to raise some external funding for such an effort. The result the CD-rom 'Söder i våra hjärtan' (Söder in our hearts) was published 1996 in conjunction with the arrangements when Stockholm was the Cultural Capital of Europe. That was the first major ALM-production in Sweden. It presented and made available some major archives as well as a large number of minor samples of the treasures hold by our cultural institutions, and it did so with a powerful yet simple to use interface. It brought the Archives to your desk and also provided you with the means to extract your own data (self service). During the development we were working closely with a class of secondary school students for the functionality specifications. We have just made a new reprint of the CD, so it still find new users after nine years. After the positive reception of the 'Söder-CD' by a variety of large user groups we have managed to produce two more with the same format apart for same simpler productions. We have moved to DVD as media to be able to include more images (photos, scanned maps and drawings and other archival materials as well as video). The last of these is 'Kungsholmen' that was released in May this year and will be used to present one of the interfaces we have developed for the presentation and provision of our demographic databases.

There are legal restrictions regarding the way in which we can disseminate information from the Roteman Archives. With people still alive and a lot of sensitive social and medical information, we are not aloud to make them accessible over the web. We have however made applications to search and download data from the Roteman files for those that we are sure is not alive.

From the start ad hoc solutions was the only way that we could serve our users different needs. This is something we still do but to a limited extent and towards special users. The knowledge we have learnt from this and the contiguous dialogue with the user has enabled us to come up with a number of solutions that can be summed up in the word 'self-service', at the same time as we have extended our services. With this self-service approach we have improved our the service we provide to our visitors at the Stockholm City Archives where we have an application, where all the digitized Roteman records so far, are individually linked.

Legal issues

Since we are dealing with information concerning individuals we always have laws, regulations and rules to follow. The legislation regarding personal integrity is an important issue on the present agenda. It has a special justification in this electronic age when we leave marks of our activities in all sorts of digital information systems. However this sometimes comes in conflict with the Archives task to provide information. We must be able to find ways where we can do the latter without getting in conflict with constraints caused by laws and regulations on integrity. Solutions to these issues are usually very specific to a research problem but by getting experience of handling this issues we hope to be able to come up with solutions that can be more generally used for our future applications for provision of individually based demographic material.

Interfaces

In order to present the dynamics in individually based demographic longitudinal data you need to find ways that can combine different pieces of information. When we have been around presenting the Roteman Archives we have tried various ways to do this. These

experiences have led us to adopt and develop a combination of text and graphics that can be interactively manipulated by the user in order to visualize the information content. Time and space are two key concepts that we use so the map becomes an important component in our scheme as well as the time-bar. This approach can be used both on the individual as well as aggregate level. (This will be presented during my presentation). We have developed this to be a part of our search and retrieval as well as analysis system. Again this can be seen as part of our strategy for self-service.

What next?

The larger the area covered by our digitizing activities the longer and more complete we can make the individual longitudinal biographies. With a large continuous geographic area mapping can be more utilized in order to fully exploit the potentials of the information in the Roteman files. The experiences of the application based on all the rotar (wards) digitized to date that we provide for our physical visitors, indicates that it will be a most appropriate base for future products. We will also extend the use of HISCO and HISCLASS-coding to facilitate statistical analysis and also use it to enable for the researcher to translate the Swedish titles to something understandable for the foreign user. Apart from performing semi automated record linkage between the entries in the Roteman Archives we will include linked data of other digitized archives. The most important of these at present is the Death Certificates with detailed notes on causes of death for each individual.

In this work we will further develop our graphics techniques for analysis, visualisation and presentation of dynamic processes. It will be an Archive that comes with a large toolset for handling the information, searching, retrieving, analysing, visualization and presentation of longitudinal data in a spatial context, 'A historical Laboratory'. As new Rotar (Wards) are digitized the database can be reloaded with even more linked individual biographies. Walking around in a 3D-world on Internet, based on yesterdays Stockholm, meeting some of the people you find in the Roteman Archives is possible for a small area right now. As a co-operation with the School authorities in Stockholm, Stockholm City Archives established a prototype web site 'The Historical laboratory' some years ago. The site was intended to provide students with digital sources from the Archives holdings. One part of the site contained a interactive 3D-world based on some blocks close to Centralstation, (The main railway station) in Stockholm in 1899. This 3D-world was partly built by the students themselves based on drawings, maps and pictures of the houses from the archives. Walking around along the streets it is possible to gather information on who lived there and what kind of activities that took place. This information can be downloaded and reused by the students. That project has been superseded by 'Stockholmskällan' a large ALM-project with the City's ALM-institutions, the City's Statistical Office and the City's School Agency. It will provide a common web site and service for primarily students and teachers in the school system, but can of course be used by anybody.

References

1 Fogelvik, S. 1989. 'The Stockholm Historical Database at Work.' History and Computing II, Manchester.

2 Fogelvik, S. 1992. 'Fiche & Chips dissemination of Historical Data: The experience of the Stockholm Historical Database.' Cashier VGI 5, Hilversum.

3 Fogelvik, S. 1995. 'The Map and the Roteman System – Geographic Information in the Roteman Archives. A useful approach?' Halbgrau Reihe zur Historischen Fachinformatik, Band A25, Göttingen.

4 Geschwind, A. – Fogelvik, S. 2000. 'The Stockholm Historical Database.' Handbook of Historical Microdata for Population Research, Minneapolis.

Towards a standard for MA programs in historical computing

The experience of Russian and CIS universities

*Irina Garskova**

The historical curriculum at Russian and CIS universities is dramatically changing under the impact of new information and communication technologies. Professional historians more and more often deal with large-scale innovations at libraries, museums, archives, publishing and multimedia houses and other repositories of the national cultural resources. New professional disciplines emerge as a response to the challenges of information technology. Advanced skills in computing are extremely needed to enable students to get a good jobs in the information society.

The specific feature of Russian system of higher education is the existence of state (federal) standard of curricula in each discipline. The standard for education in humanities includes the 'hard sciences and mathematics' division with such obligatory courses as mathematics and informatics (approx. 150 hours in 4 terms). Thus, all history students during the first or second year of education learn courses, which deal with the basic understanding and skills in practical use of computers: familiarity with hardware and operating systems, additional utilities, the Windows system, text processing, Internet access. Such 'computer literacy' courses often deal with technical aspects of computing and taught by computer scientists. Experience learns that pure computer literacy courses have no effect, unless students see links to their sphere of interest.

We absolutely agree that it is pointless to teach history students computer science unless it is not directly related to their domain of expertise and that courses on historical informatics will remain a transient phenomenon unless they include an thorough understanding of the substantial requirements of historical research on the one hand and computer science concepts on the other – opinion, which have been expressed by members (M.Thaller. K.Smedt, T.Orlandi, etc.) of the working group in 'Formal methods in the Humanities and their teaching' of the network project on Advanced Computing in the Humanities (ACO*HUM).

The best solution is to teach obligatory courses on informatics providing students with training in the methodology and techniques of computer-based historical study (unfortunately, some universities face difficulties in developing such model because of the lack of human resources and literature on historical computing).

The majority of Russian and CIS universities, well equipped with computers and competent personnel, are able to provide dedicated courses in historical informatics at the history faculties. Such 'basic' courses tend to present a lot of tools and problems to be solved by them, to show which tools should be applied to specific kind of problems (typology, dynamics, etc.), to describe specialized non-commercial software, which have been elaborated for historians. Teaching is based on the set of examples, fragments of primary sources, and control tasks extracted from some historical research in the field of quantitative history and historical informatics. The course usually includes such themes as the range of concepts, methods, techniques and issues which predominate in historical computing; fundamentals of text processing and analysis (including scanning and OCR); structured data processing (spreadsheets, databases, statistical methods); image processing; digital resources and information systems (including Internet). Each theme could be learned later in more details either in the specialization framework

*Moscow State University

or in an optional course.

Thus, there are two ways to learn historical informatics after the 'basic' course. If a university does not offer the dedicated specialization, highly motivated students could acquire advanced competence through optional courses; if the specialization exists, he (she) could choose to learn historical informatics during next three years.

Some CIS universities have already established historical computing (historical informatics) as a specialism and included MA in historical computing in the general curriculum. Different models of MA programs in historical computing have recently been opened at the Byelorussian State University (1995), Mordovian State University (1996), Stavropol State University (1998), Russian State University for Humanities (2003) and some other CIS universities. The significant event of the last year was trasformation of the MSU Laboratory of historical information science (the leader of historical computing Russia and former Soviet Union) to the MSU Department of historical information science. Although it is good to have a variety between programs at different universities, the new department supposed to discuss a possible standard for MA program.

Analysis of different universities' programs in the field of historical informatics revealed the following preliminary list of core curricula components:

– General introduction: overview of the subject (theoretical and historiographical aspects), introducing to the range of concepts, methods, techniques and issues which predominate in historical computing;
– Fundamentals of information technology for historians;
– Advanced information and communication technology for historians (including expert and artificial intelligence systems).
– Data bases in historical research (concepts, design and manipulation);
– Quantitative methods in historical research;
– Data analysis of historical numerical data using statistical software packages (descriptive and inductive statistics);
– Text management and analysis (including textual data base management systems and principles of the content-analysis);
– Computer networks (data communication and exchange with Intranet/Internet);
– Digital resources for historians (on CDs and Internet): creation and use;
– Legal aspects of electronic editions of historical sources;
– Image processing (fundamentals of the use of graphics and graphical packages);
– Multimedia in historical research and education;
– Electronic documents and archives;
– Automation in archives, libraries and museums (data bases and information systems, electronic catalogues);
– "Historian-oriented" algorithms and software (such programs as well-known Kleio, FuzzyClass for fussy classification, information system ProSys for prosopography, Manuscript for processing of old Russian texts, a set of program for computer-assisted simulation, etc.);
– Computer-based simulation of historical processes;
– Geographic information systems in historical research;
– Computer-aided methods and information technology applications in socio-economic history and historical demography (optional course).
– Fundamentals of algorithmic and programming knowledge;
– Procedural and object-oriented programming;
– Introduction to proposition and predicative logic;
– Introduction to probability theory and mathematical statistics.

There exist many differences in the content of curricula between universities, although the most advanced universities form more homogeneous group and their curricula are rather similar. Nevertheless, that group also demonstrates some peculiarities. For instance, the Moscow State University program includes deeper study of application of computer based statistical methods and simulation in historical research, especially in socio-economic history as essential part of the specialization. The important part of Byelorussian State University program is computer-assisted instruction, Moscow State University of Humanities program pays more attention to application of information technologies in archives.

To overview Russian and CIS programs on historical informatics, we have reduced information in table that shows blocks of courses (modules) which are more often included in curricula in different universities, and the variety of courses falling into these blocks:

Modules	Available in curricula	Number of courses
Mathematics for historians	++	4
Computing and information technology for historians	+++	5
Programming	++	4
Text management and analysis	++	6
Analysis of structured information	++	5
Multimedia	++	6
Local networks and Internet	+++	7
Quantitative methods and simulation in historical research	++	4
Electronic documents and archives	++	5
Automation in documentary information	+	1
Information technology in archives, libraries and museums	++	4
Information technology in higher education	+	2
Optional courses	++	6

+++ A module is available in all curricula
++ Available in a half of curricula or more
+ Available in less than a half of curricula

The specialization exists at advanced levels of education (MA or – Russian equivalent of MA – 'specialist', a degree that is assigned after 5 years of university education); it takes approx. 800 hours in 6 terms or 1000 hours in 10 terms (Moscow State University of Humanities). Some universities offer additional interesting activities for students: summer training courses at archives, museums and libraries, participation in the research projects or projects devoted to the electronic resources of teaching programs creation, etc.

The program of specialization culminates in a dissertation project, devoted to investigation of an historical topic. Dissertation should explain methodological and technological aspects of the research and contain an application of computing tools and information technologies appropriate for the research topic. Results and their interpretation should demonstrate the significant role of historical informatics methodology in posing and solving historical problem.

The overview of students' diplomas and MA dissertations shows that dominant topics are: applications of quantitative methods (in particular, multivariate statistical analysis) in social and economic history and historical demography; producing electronic resources for historical research (for instance, databases creation and analysis in prosopography) and education (*e.g.*, teaching resources for distance learning); producing information tools and computer programs for information systems in archives, libraries and museums.

The integration of multimedia resources and teaching programs into 'classical' model of education in humanities enables new opportunities to teachers and students in using electronic encyclopedias, electronic editions of historical sources, teaching programs, electronic textbooks, electronic libraries, Web-portals, devoted to education, etc. On the other hand, computer is becoming an essential tool for the study in all historical subdisciplines. The education in the field of historical informatics, interdisciplinary by nature, provides critical analytical thinking and practical skills with computers and software. It means that students trained in the specialization acquire a highly transferable and universal knowledge that enable them to be attractive to employers.

Indigenous peoples of the North-western Siberia: Ethnohistorical mapping[1]

*Elena Glavatskaya**

This paper presents the computer and other methods and sources used to map the ethnic history of the Khanty and Mansi. Since their remote location and history is not well known, I need first to present an introductory overview. This is based on my articles on the Khanty and Mansi ethnohistory as well as my dissertation 'Politica russkogo pravitelstva v otnoshenii yasachnogo naseleniay severa zapadnoi Sibiri v 17 v.' [Russian state politics towards the indigenous peoples of the Northwestern Siberia in the 17th century] (1992) and book (forthcoming 2005) 'Religioznye traditcii Khantov v XVII-XX vv.' [Religious traditions among the Khanty. 17th-20th centuries]. Three sample maps are attached, more will be presented during the conference.[2]

Introduction

The two main indigenous people of the Northwestern Siberia are the Khanty and the Mansi (formerly also known as the Ostyaki and Voguly). They belong to the Ugrian branch of the Fenno-Ugric linguistic family, mainly dwelling in the basin of the Ob' river and also known among ethnologists as Ob'-Ugrians. Both minorities belong to the twenty-six widely dispersed small indigenous peoples of Northern Russia. According to the 1989 census they numbered 22,283 and 8459 persons respectively. The great majority of the Khanty and the Mansi still live in *Khanty-Mansieskii Autonomnyi Okrug and Tjumenskaia oblast'*, while some Khanty groups live in *Yamalo-Nenetskii Autonomnyi Okrug and Tomskaya oblast'* and some groups of the Mansi in *Sverdlovskaya* and *Permskaya oblast'*. All in all the contemporary territory, settled by Ob-Ugrians approximately covers a vast area of 900000 sq km.

Both the Khanty and the Mansi can be divided into different groups distinguished by their means of subsistence, language and culture. Their economy was largely based on fishing, hunting and gathering, supplemented by reindeer herding in the north, and agriculture including cattle breeding in the south. The southern Khanty as well as the western and southern Mansi were incorporated into Russian society both economically and culturally by the middle of the 20th century, and no longer live a traditional way of life. However, the northern and eastern Khanty and the northern Mansi have managed to maintain large parts of their traditional life, although they have been considerably affected since the 1960s by the intensive industrial development in the region.

Short ethnohistorical survey

By the beginning of Russian colonization in Siberia at the end of the 16th century, there were several Khanty and Mansi principalities of varying sizes in Northwestern Siberia. The 16th century was a period of political consolidation when the major principalities struggled against each other for political superiority. Russian conquest and the system of rule established in Siberia affected the social and political structures of the Ob-Ugrian societies and hindered their autonomous development. A few Khanty and Mansi nobles died during wars and uprisings against the Russian power, some were taken as hostages to Russian towns while others had to accept subordination to Russian supremacy in Siberia. Since the beginning of Russian colonization, the indigenous peoples were proclaimed object to the Russian Tzar's taxation and were forced to pay '*yasak*' – a fur tax. The average value of the *yasak* at the turn

*Urals State University, Lenina Ave 51. Ekaterinburg, 620083, Russia, glav@tehnec.ru

1 The research was supported by RGNF (Russian Humanitarian Scientific Foundation), grant No. 04-01-00138a

2The attached maps have legend and names in Cyrillic that will be transferred into Latin for the presentation during the conference.

of the 17th century was 5-12 sables per person per year, roughly corresponding to the price of a cow. Being economically interested in and dependent on valuable furs, the Russian state promoted a policy to conserve the social structure and traditional way of life of the Khanty and Mansi. Because of this, some Khanty and Mansi groups managed to maintain their traditional patterns of land-use until the 20th century almost unchanged. Even under Soviet rule, some of them did not experience significant social change, because their regions lay beyond the main Siberian crossroads and were always considered to be inaccessible both during winter and summer. Even if the politics of establishing the 'socialist' form of land-use, '*kolkhosy*' or collective farms was quite harmful for the Khanty and Mansi in general, it did not dramatically alter the life of some remote northern and eastern communities in the long run.

Perhaps the most harmful for traditional Ob-Ugrians' land-use and life was the Soviet policy directed toward settling them in specific, huge national settlements in the late 1940s and 1950s. The main idea was to create better living conditions, access to medical care and education, in order to promote 'advanced', 'progressive' and 'civilized' forms of life. Unfortunately, this policy brought more disadvantages than advantages to the Khanty and the Mansi. On the one hand, concentrating large numbers of people in a few places brought a decline in reindeer populations and devastated the hunting, fishing and gathering estates in those vicinities. On the other hand, this policy also increased the dependency of the ethnic minorities on the Soviet system for goods distribution and social support. Eventually this policy led to a reduction of the Khanty and Mansi ethnic territories. The loss of necessary resources for a traditional way of life also increased the process of assimilation or those who were forcibly resettled. Nevertheless, those who were not assembled in settlements, successfully maintained their traditional patterns of life.

This situation changed dramatically once again in the late 1960s with the first discoveries of petroleum deposits in the ethnic territories of the Khanty and the Mansi. In the late 1980s the Ministry of Energy and Oil Industry seized huge territories of Siberian indigenous peoples for oil production. The Soviet government, however, together with local authorities, always had to make certain provisions, such as adjustments to state subsidies, guarantees of employment, and organization of systems for purchasing their hunting, fishing, and gathering production in order to alleviate some of the economic and social stress of the destructive industrial intervention upon traditional life. Thus, even if the basis for a traditional way of life had been continually eroded, the Soviet system of social security and support directed to the most vulnerable groups and national minorities gave the Ob-Ugrians at least some guarantees for a minimum level of well being. So from this perspective the collapse of the Soviet Union has been disastrous for Northwestern Siberia and its indigenous peoples. By the late 1980s, the consequences of the oil boom became even more devastating economically, socially, culturally and ecologically for the Khanty and Mansi lands. A detailed analysis of the impact of oil production upon Khanty traditional land-use was made by Andrew Wiget[3]. All of these factors constitute a direct threat to the very existence of the Khanty estates and the resources necessary to sustain a traditional way of life.

In the Mansi case the late 1980s was marked by the collapse of the developed GULag (forced labor camps) chain in the region together with the Soviet system. Both formerly made their painful impact on the Mansi lands and their ethnic and religious heritage, but also brought some development of the infrastructure and social support for the people in the region. After the collapse of that system, the Mansi were left alone to deal with numerous problems which used to be the state's duties before. The solution was found like in the case of the Khanty partly in returning back to their traditional values, which helped their far ancestors to survive.

The system crises of the early 1990s created a situation in which the Khanty and the Mansi found themselves endangered both as ethnic communities and individuals. As the Khanty scholar Tatiana Moldanova pointed out '...a person has been left face to face with an impending danger, prepared neither physically nor psychologically to resist it. A set of protective mechanisms, transmitted by the ancestors, turned out to be inadequate to protect against perils created by a new situation and reality.'[4]

3 Wiget A. *Chernyi Sneg: Neft' i Vostochnye Khanty*. [Black Snow: Oil and the Eastern Khanty. *In Ocherki Istorii Traditcionnogo Zemlepol'sovania Khantov Materialy k Atlasu)*. [Essays on History of traditional Land-Use of the Khanty (Materials for the Atlas). Ekaterinburg, 2002 'Tezis'. [2-ed.]. P. 211-222.

4 Moldanova T. *Rynok dlia aborigenov* [Market for the indigenous people]. In. '*Rissiiskaia Federatcia'* [Russian Federation]. 1996. Issue No. 19. P.7.

Mapping approaches

The project to be presented at the AHC conference is devoted to creating an Atlas of the history of the Khanty and Mansi in the late 16th to 20th century, while they were exposed to a strong foreign political, economic and cultural influence. The mapping of ethnohistorical changes helps visualize the historical process and facilitates prognoses on future developments of ethnoreligious interaction in the region.

The main effort in the project was not put into mapping the general history of the region itself, which is the traditional approach when making atlases, but on the history of the indigenous peoples and how the historical events affected their way of life. So the full detail, which used to be ignored before would be mapped. In order to achieve this aim information was extracted from different types of sources: state and local laws, official reports, correspondence between the Siberian officials, missionaries', travelers', and exiles' diaries, complaints made by indigenous people, folklore, tax books, archaeological investigation reports, historical maps etc.

As a result, three main clusters of ethnohistorical factors were distinguished to be mapped: first the Khanty and Mansi traditions, second events which caused the change of traditions, and third innovations that were caused by either the state politics or intensive cultural interactions. The main problem arising at that stage was to place all the ethnohistorical factors not only in time but most importantly in proper space. Most of the sources on the ethnohistory of the Khanty and Mansi lack correct and detailed geographical information, and handmade maps lack much detail. Thus, even knowing verbally described location of the given site (say left bank or a mouth of some small tributary to some river) it was impossible to find the proper spot on the map. So the task of ethnohistorical mapping required additional methods.

Computer resources and methods

In order to solve the problems of ethnohistorical mapping some necessary resources were made available by our computer systems. First, a digital map of the *Khanty-mansiesk autonomous okrug* became the base for further research. Its main advantage is the comprehensive detailed information on hydro resources in the area. That helped to locate some of the sites, but not all of them. Sometimes all that was known about a certain site was its location relative to a certain settlement that existed in a previous century and was not marked on the contemporary digital map. So in addition to the digital map those drawn by explorers in the 19th century were used. Together these two resources provide the opportunity to locate most given sites in a proper place.

To map the different events and factors a system of symbols was elaborated. However, in cases when similar events took place in the same area several times within a very short period (like itinerary of the missionaries, baptism, clashes, passive resistance of the indigenous peoples, destruction of their sacred places, etc.) additional means were needed. In such cases computer facilities provided a rich opportunity to use multiple colors to distinguish one campaign from another.

For to the tasks of this project and due to the available basis for electronic mapping the CorelDRAW program was used. The process of mapping itself consisted of the following stages:

1. To make a chronological sequence of basic, digital historical (administrative) maps of the area from contemporary digital and historical maps.[5]
2. To distinguish important ethnohistorical events on the basis of searches in different types of sources.
3. To find their geographical locations
4. To make connections to other simultaneous factors and events
5. To elaborate certain symbols
6. To draw a map on the printed copy of a digital one
7. To computerize the result[5]

Preliminary results and perspectives

As a result a set of digitized ethnohistorical maps were prepared: 'Sacred places of the Khanty and the Mansi'; 'Administrative map of the Northwestern Siberia in the 17th century'; 'Religious situation in the late 16th-17th century'; 'Anti-Russian riots in the 17th century'; ' Russian Otrthodox Church Missionary campaigns in the beginning of the 18th century'; 'Administrative map of the Northwestern Siberia in the 18th century'; 'Spreading of Christianity in the 18th century'; 'Administrative map of the Northwestern Siberia in the 19th – beginning of the 20th century'; 'Spreading of Christianity in the 19th – beginning of the 20th century'; 'Development

[5] This stage of work was conducted together with/ or by a computer expert Svetlana Tcemencova

of the school education for the Khanty and the Mansi in the 19th – beginning of the 20th century'; 'Exploration of the Khanty and Mansi territories by Finnish and Hungarian scholars in the 19th – beginning of the 20th century'; 'Establishment of Soviet power in the Khanty and Mansi lands'; 'Anti-bolsheviks riots in the Khanty and Mansi lands'; 'Collectivization and opposition to it'; 'Khanty and Mansi shamanism in early 20th century'; 'Religious changes among the Khanty and Mansi in the 20th century'.

The next steps planned are to work from the contemporary GIS maps in order to transfer the ethnohistorical Atlas of the Khanty and the Mansi into software such as Mapinfo professional, combining this with pictures, video and audio into a multimedia presentation.

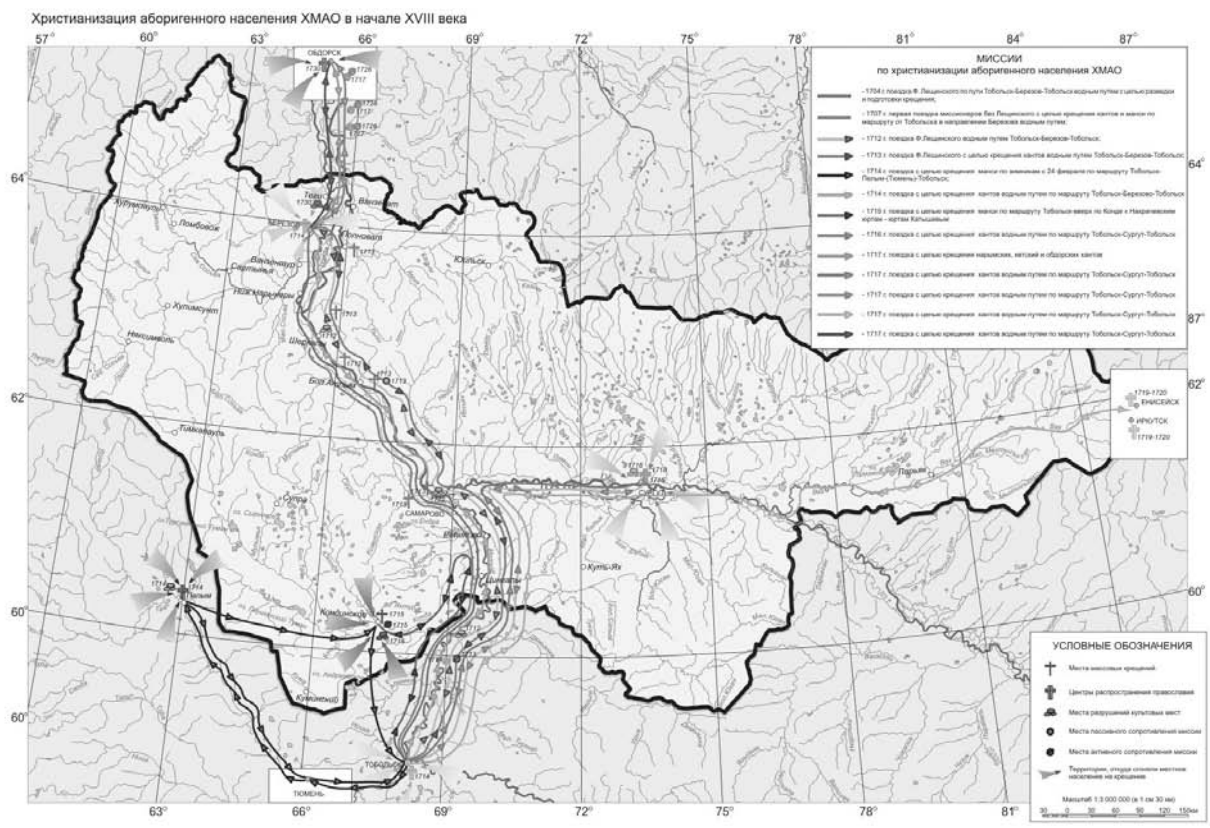


Figure 1. Russian Orthodox Church Missionary activity in the Khanty and Mansi lands in the early 18th century

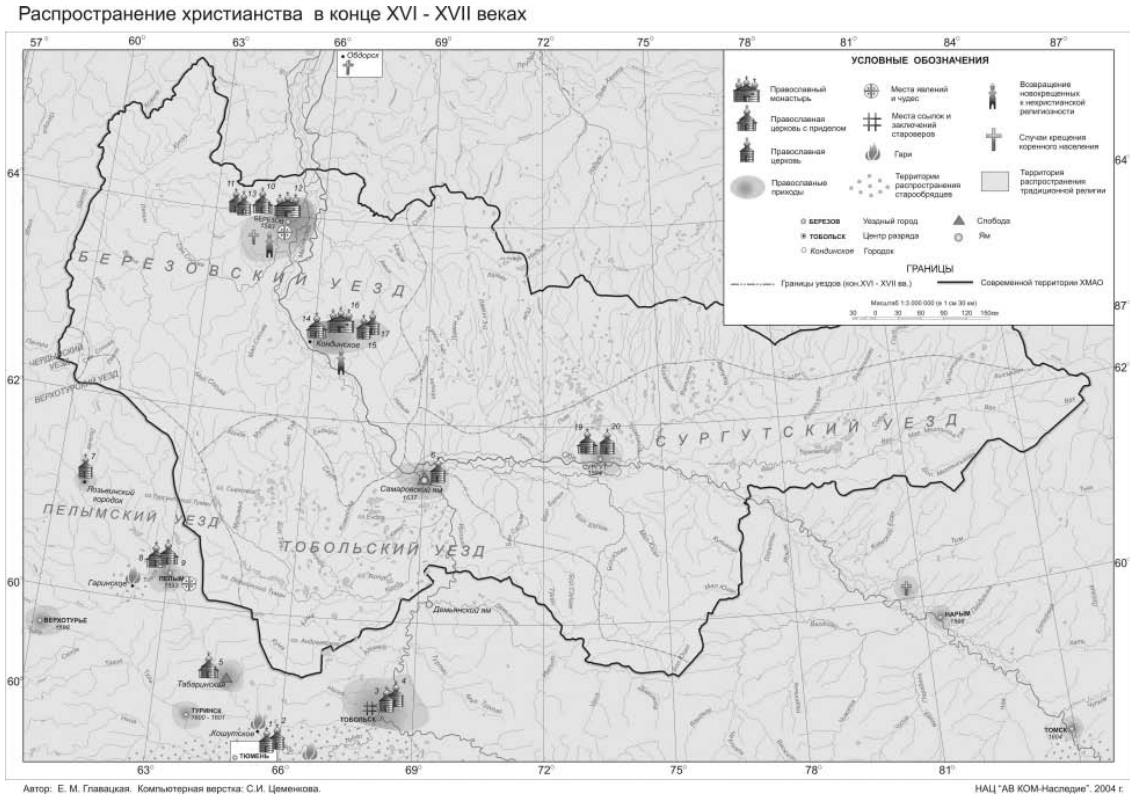


Figure 2. Spreading of Christianity in the Nothwestern Siberia in the late 16th-17th centuries Churches

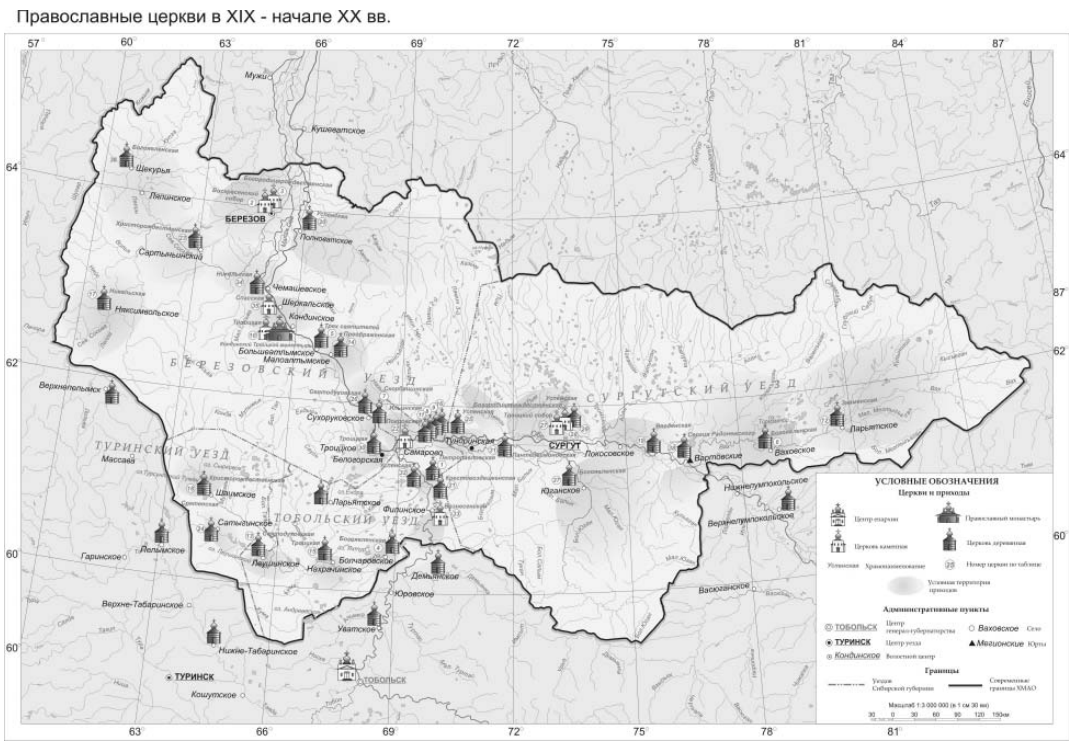


Figure 3. Churches

Creating analytic results from historical GIS

*Ian Gregory**

National historical GISs (NHGISs) are expensive and time consuming to build. They have attracted a lot of funding in many countries and there has been a considerable amount of hype about their potential. In spite of this, knowledge about what an NHGIS can actually offer to historical scholarship is surprisingly limited. In many cases people believe that the main output from an NHGIS will be maps. Maps tell us very little and usually lead to more questions than answers and the main strength of GIS is not as a mapping tool, but as a tool for geographical analysis, or put more broadly, a tool that allows us to analyse historical data in ways that explicitly includes the geographical dimensions of the issues under study.

GIS is a far from perfect tool for doing this. It is based around precise co-ordinates that allow no ambiguity and tends to see 'geography' as the straight line distance between places. Much of the data contained in NHGISs is also very limited. In particular, it tends to be based on modifiable areal units such as districts, counties and municipalities. These tend to sub-divide the population using boundaries that have no actual meaning on the ground. An additional and frequently overlooked problem with using modifiable areal units is that they may have vastly different populations, for example in England & Wales in 1911 registration districts, the unit used in this analysis had two districts with populations of less than 2,500 (the Scilly Islands off of Cornwall, and Reeth in the North Riding of Yorkshire) and two with population of more than 250,000 (West Ham in Essex and West Derby in Lancashire). Given these huge variations in population and the fact that these units are almost completely arbitrary subdivisions of the country it makes little or no sense to calculate the average rate for a district with the Scilly Isles having the same weight as West Ham.

Never-the-less, the use of NHGISs and zone-based quantitative data does offer up an exciting potential to analyse the geographies of the past and one that must be realised if we are to justify the investment in such systems. In previous work I have described spatial statistical approaches such as the use of areal interpolation and geographically weighted regression to explore population change during and after the Great Irish Famine of the late 1840s. Spatial statistical approaches are basically statistical approaches with additional spatial characteristics. In this paper I want to change this emphasis and explore possible approaches to analysing long-term geographical change that emphasise the spatial rather than the statistical. The paper concentrates on infant mortality from the 1850s to the 1900s. At the moment it is still work in progress rather than a finished article and in particular I would emphasise that I am more interested in presenting the techniques and possibilities rather than the substantive results and additional knowledge about changing infant mortality over this period.

Over the period infant mortality in England & Wales started with 153 deaths/1,000 live births in the 1850s but declined slowly throughout the period to reach 128 deaths/1,000 live births in the 1900s. The 1890s were an exception to this decline with notable higher rates than the surrounding decades. One possible explanation for this is that this decade had a number of hot summers that encouraged the spread of infectious diseases such as diarrhoea that were major killers of infants at that time. Mapping the pattern suggests that the spatial pattern of inequality in infant mortality varies over the period, however, given that there are six maps, one from each decade, and each map contains over 600 districts exploring how the pattern has changed over time is very difficult. For this reason analytic approaches are required to explore the geographical divides in infant mortality.

Centre for Data Digitisation and Analysis, Queen's University, Belfast

Firstly we explore whether there is an urban-rural divide. This does not require a GIS, it can be determined using basic exploratory statistical procedures. Defining 'rural' and 'urban' is problematic. Here I have gone for a simple measure based on population density. All of the population densities for every decade were brought together and the divided into eight classes using nested-means. In other words, the average population density was taken as the middle subdivision to create two classes, the average of those above the mean and those below the mean was used to create four classes, and this was repeated again to create eight. This gives a crude hierarchy of the degree of urbanisation of each district that allows districts to become more (or less) urban at a time of rapid urban growth. In this case equal weight have been given to districts as we are attempting to define the districts themselves rather than use the districts to make a statement about broader trends.

Once we had divided district in this way the infant mortality rate for each type of district could be calculated and graphed in figure 1. The slide shows the changing average infant mortality rate for each type of district expressed as a percentage of the national average infant mortality rate for that decade. It can be seen that the two most rural classes in particular start with rates well below the national average and improve significantly from here. At the other extreme classes 8 and 6 both start with above average rates and stay the same or get worse. The class between these, class 7, on the other hand does show some improvement from well above average to slightly below. Of the middle three classes, 3 and 5 show some improvement while 4 remains roughly constant. This contradicts much of the accepted logic of infant mortality decline in this period which states that this did not start until the 1880s. Clearly for the more rural areas this was under way by the 1860s.

Let us now turn to the north-south divide or core-periphery divide as it is more correctly known. Although much has been written about this divide it has never been well described. Approaches to its definition usually either involve drawing an arbitrary line across the country usually form the Bristol Channel in the west to the Wash or the Humber in the east and declaring those areas south of the line to be 'core' and those north of the line to be 'peripheral.' An alternate approach is to take a sample of places in the north and compare these with a sample of places in the south. Neither of these is satisfactory as the lines on the map between core and periphery tend to be largely arbitrary while the sample approach leads to the suspicion of

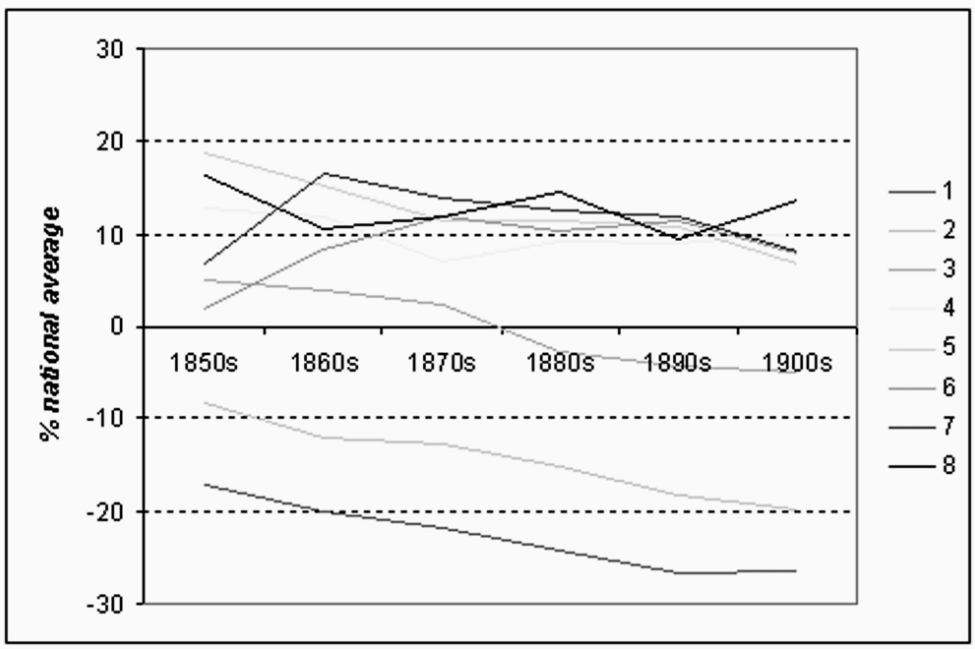


Figure 1. Changing infant mortality in England and Wales by urbanity, 1851-1911. Class 1 are the most rural, class 8 the most urban

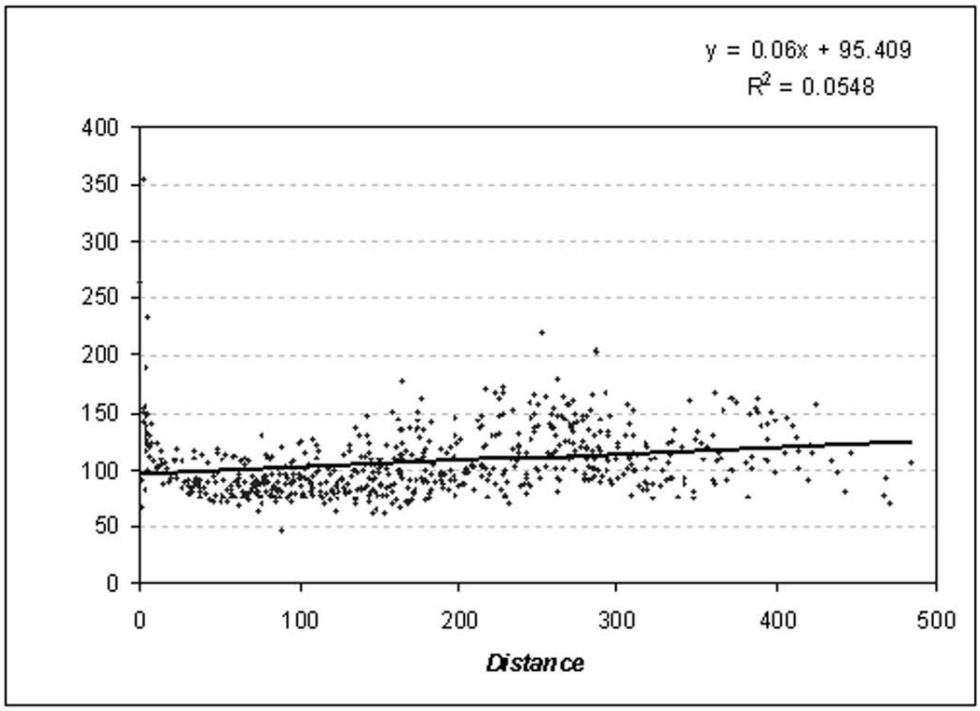


Figure 2. Infant mortality rates with distance from London, 1911

bias in selecting areas to compare.

In this study we resort to basic principals and use GIS to implement them. If we claim that there is a core-periphery divide then we are arguing that in the core, ie London, there should be low rates and as we move away from London into the periphery these rates should rise increasingly. This applies to rates of anything but in this case we are looking at infant mortality. Using GIS we can test this. We define the core as a point in central London, for convenience Nelson's Column has been used. For each district we then take the geographical centroid and measure the distance from the centroid to central London. This gives a statistical measure of how peripheral each district is. Using this we can create a scatter graph of infant mortality rates to distance from central London. From this scatter-plot we can create a trend line and measure the slope, intercept and statistical significance of the trend line using OLS regression. The intercept gives the infant mortality rate in central London predicted by the trend-line, the slope gives how quickly this rises from London, while the R^2 value gives a measure of how much of the total variation can be 'explained' by the distance measure.

Figure 2 shows the pattern for the 1900s. The intercept is 95.4 and the slope is 0.06 which tells us that in central London we would predict an infant mortality rate of 95.4 deaths/1,000 live births and that this will rise at 6 deaths/1,000 live births for every 100kms we move away from London. The R^2 is 5.5%, quite a high value such crude data in social science. The pattern is statistically significant at the 1% level showing that it is highly unlikely to have occurred at random.

Table 1. Regression results for the core-periphery divide, 1851-1911

	Slope (100kms)	Intercept	R2 (%)	F	Signif
1900s	6.00	95.41	5.5	36.64	1%
1890s	4.88	118.5	3.5	22.57	1%
1880s	3.04	115.31	1.7	9.59	1%
1870s	2.72	124.14	1.4	8.75	1%
1860s	1.16	132.3	0.2	1.35	No
1850s	-3.20	143.14	1.7	10.71	1%

Table 1 shows the regression equations for all six decades. The results are interesting: as we move back from the 1900s the slope and the R^2 both decrease steadily although there remains a statistically sig-

nificant relationship between distance from central London and infant mortality. By the 1860s there is no statistically significant relationship and, perhaps most surprisingly, in the 1850s there is actually a statistically significant negative relationship implying that in this decade there was actually a south-north rather than north-south divide. This suggests that from the 1870s on there was an increasing core-periphery divide, however, there are potential problems with this analysis. Regression analysis uses various assumptions that are often violated particularly with spatial data. In this case there is clear dependency in the independent variable (distance from London) as well as in the dependant variable (infant mortality). There is also spatial autocorrelation in the residuals. On a more general note, an aim of this paper is to use spatial rather than statistical approaches.

How else can we explore whether there is a core-periphery divide? A simple GIS-based approach has been to sub-divide the country buffers 25km wide based on their distance from London. The number of infant deaths, births and the total population from the centroids falling in these buffers has been summed and the infant mortality rate for each buffer calculated. Doing this shows that in the 1900s there clearly was a core-periphery divide. Apart from the inner circle that includes London there are then two circles in the best decile followed by three in the next best decile. The nearest worst decile circle runs through Leeds and no buffers north of Coventry are in the best three deciles other than the three extreme northern buffers all of which have only a very small number of districts with small populations. The 1850s shows a more complex pattern with London being surrounded by good areas, then the area from Birmingham to Liverpool/Manchester/Leeds being bad, but further from this the areas are again generally good.

Figure 3. A growing core-periphery divide.

Relative change in infant mortality between the 1850s and 1900s by distance from London. The map uses 25km buffers. Buffers that have got relatively worse over the period are shaded red, those that have improved are shaded green. Yellow shows no change. Darker shades indicate a larger change.

This suggests, therefore, that over the period from the 1850s to the 1900s while overall inequality in infant mortality did not seem to increase much, there was an increasing polarisation of the country with the area around the core having increasingly low rates with high rates increasingly polarised in the peripheral areas. Figure 3 summarises this by comparing the decile that each buffer lay in in the 1850s with the buffer that it lay in the 1900s. Those that moved into lower rate buffers are shaded green while those that moved into higher rates are shaded red. The pattern is unequivocal: the peripheral areas got worse while the core ones got better. If there is a dividing line then Nottingham and Bristol are on the right side of it while Cardiff and Sheffield are on the wrong side.

We have established so far that there appears to be a rural-urban divide that grew over the period as a result of improvements in rural areas and a core-periphery divide that became increasingly established from the 1870s onward. A major problem with this analysis is how do we separate the rural-urban divide from the core-periphery divide given that it may well be that the northern industrial areas centred on Liverpool, Manchester, Leeds, Sheffield and Newcastle may be driving this pattern. To attempt to pull these factors apart we draw on the ideas of Skinner *et al* (2000) and create a hierarchal settlement matrix. We already have and eight way urban hierarchy based on population density. We can also create an eight-way periphery hierarchy based on distance from London by dividing the country into bands 50km wide. Thus band 1 is all districts

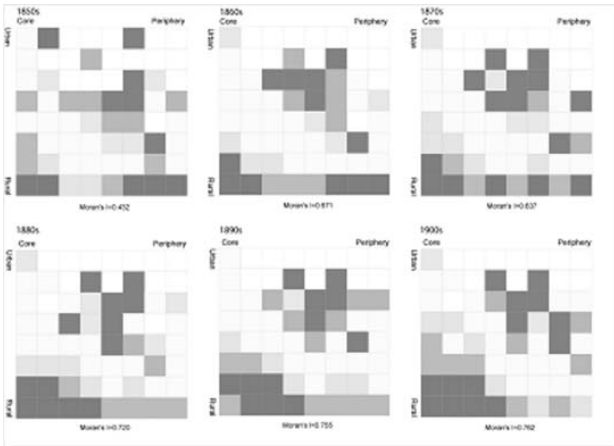


Figure 4. Hierarchical settlement matrices for infant mortality, 1851-1911

whose centroid lies within 50km of central London while band 7 is districts whose centroid lies between 300km and 350km from central London and band 8 is all those districts lying more than 350km from London. Square 1,1 is the core urban areas and includes much of inner London, square 8,8 is peripheral rural and includes much of rural Northumberland, the Lake District, Anglesey and parts of Cornwall. Square 1,8 are very rural areas near to London while square 8,1 would be very dense urban areas in the periphery of which there are no values. To create time-series for districts lying in some of these squares we have had to aggregate some squares together such that they have at least one observation for each decade.

Core urban areas are in the top left of each matrix, peripheral rural ones in the bottom left. The legend sub-divides the population into deciles with even numbers of people. The best three deciles are shaded in increasingly dark shades of green, the worst three in increasingly dark shades of red.

In figure 4 the infant mortality values for each cell in the matrix have been calculated for each decade. The pattern shows clearly that the urban peripheral areas have the highest rates quite consistently. These start in around band four, which contains Bristol, Birmingham and Nottingham. The rural-urban pattern is more complicated with the most rural band having among the lowest rates in the earliest decades but becoming worse in the later decades.

This work shows how we can use the analytic power of GIS to provide new insights into long-established research questions. Maps are only a small part of providing the answers. For the most part we want to use the analytic power of GIS and its ability to handle space effectively to devise new and powerful approaches to historical geographical analysis.

Occupational migration in Albania in the beginning of the 20th century

Siegfried Gruber

Migration in Albania

Recent studies have dealt with the impact of recent migration on the development in Albania (Kaser, Pichler, Schwandner 2001; King, Mai, Schwandner-Sievers 2005). Migration turns out to be the most important experience in Albania besides the change of the political, social, and economic system. Migration influences demography, economy and every-day-life in Albania (other European regions as well, but not in such a massive way) dramatically.

Migration has been recorded for Albanians already before and some scholars have even portrayed Albanians as being nomadic throughout the Middle Ages and the Early Modern Period. But elaborated agriculture existed in Albania before the beginning of the Ottoman conquest and the theory of the nomadic Albanians cannot be proved (Shkurti, Kaser 1997). The inclusion of Albania into the Ottoman Empire caused migrations to Italy, Greece and Kosovo. Migrant workers sought employment as far as in Istanbul and in Egypt and overseas emigration also began at the end of the 19th century (Kora 2000). There were also smaller waves of in-migration: Muslims leaving territories lost by the Ottoman Empire to Christian states in the 19th and 20th centuries.

Another type of migration was internal migration, mostly because of marriage patterns, where women had to come from other villages. We should also not forget seasonal migrations connected with animal husbandry. There was a periodical moving of the flock of sheep and goats from the winter pastures in the lowlands to the summer pastures in the highlands and back to the winter pastures. Such seasonal migrations could include many people in the case of large livestock holdings. In the case of the tribal territories in the north of Albania these migrations were within the defined tribal territory and therefore not far reaching. In other parts of the country such migrations could be far reaching (Kaser 1992: 55-78). Such migrations could be a starting point for other economic activities as well.

We should also not forget about migrations caused by wars and in the case of Albania at the beginning of the 20th century the effects of the Balkan Wars and WWI. The Balkan Wars ended the status of Albania as a part of the Ottoman Empire and opened the new phase of political and military attempts to partition Albania among its neighbouring expanding nation states and other attempts to create an independent Albanian state. Another consequence of the reduction of the Ottoman Empire in Europe to a small region around Istanbul was that former internal migration within the Ottoman Empire became now migration to foreign states.

Urbanisation and occupational migration

Urbanisation was only a minor phenomenon in Albania 100 years ago but still there was migration into the larger settlements. The urban population amounted to 12 percent of the whole population (Seiner 1922: 6) and the proportion of city dwellers born outside of the respective city reaches from 7.5 percent in Tirana to 33.9 percent in Durrës.

Historical migration studies for Southeast Europe lack very often a source for quantitative analysis on a larger scale. V. Duka does not deal with the question of the place of origin of the urban population in her study about urban life in Albania at the beginning of the 20th century (Duka 1997). N. Todorov gives only limited information about the origin of the urban population of Bulgarian cities in the 19th century (Todorov 1983).

In contrast to this situation A. Steidl published a study about the mobility of the people engaged in handicrafts in Vienna in the 18th and 19th century. She also deals with the place of birth of these immigrants to Vienna (Steidl 2003: 156-245). Mobility of young people was high in pre-industrial Europe and an integral part of life at this stage (Mitterauer 1985; Ehmer

1988; Schofield 1987). But the situation in Southeast Europe was different since artisans were organised in a different manner and the pattern of young people circulating between different employers (agricultural or non-agricultural) did not exist there. Within the Ottoman Empire artisans were organised within an 'esnaf' and the most important trades were reserved for Muslims (Gross 1987: 13). Towards the end of the 19th century industrial products, mostly from abroad, began to replace the artisans in the mass production of goods. Cheap labour costs and high transportation costs of foreign goods hindered a faster growth of the import trade (Okyar 1987: 191).

Migrant workers were a common phenomenon in Southeast Europe. There was a region of origin of migrant workers which in the late 19th century extended from the western borderlands of Bulgaria and the adjacent regions of southeast Serbia to the west and south through Old Serbia (Kosovo), the Macedonian vilayets and the Pindus (Palairet 1987: 23). Research about them has mostly excluded villages in Albania which are in fact neighbouring this belt. It was practiced especially in mountainous regions and formed the most important source of income besides animal husbandry. Migrant workers were mostly involved in handicraft and less often in agricultural work. Especially the construction trade was employing migrant workers. These migrant workers were predominantly male and sometimes they returned only after years to their native villages. They brought money to their families and villages and were able to buy land and construct houses. Many of them went to the larger cities within the Ottoman Empire where there were already Albanian colonies.

A new source for migration studies

There is not much known about the quantitative importance of the different types of migration at the beginning of the 20th century in Albania. A source forgotten for a long time could now bring new information to migration studies about Albania at that time. This source is the population census done by the Austro-Hungarian army during WWI[1]. The Austro-Hungarian army occupied the majority of the territory of the newly created independent Albanian state and established a new administration. The 1918 Albanian population census counted 524,217 persons who lived in about 1,800 villages, towns and cities on the territory which was administrated by Austria-Hungary during WWI (Seiner 1922: 5).

The research project 'The 1918 Albanian Population Census: Data Entry and Basic Analyses'[2] funded by the Austrian Science Fund (Fonds zur Förderung der Wissenschaftlichen Forschung) and lasting from 2000 to 2003 aimed to bring the data into a machine-readable form. This database is also an excellent source for migration studies since it contains for each person the place of birth and the place of residence. The data is still on the individual level which allows much more research than aggregate data on the village level. The researcher is able to aggregate data in a way he/she wants to do it and is not bound to the categories of already aggregated data (Hall, McCaa, Thorvaldsen 2000: 9). This also enables the researcher to combine different variables on the individual level for research purposes. Up to now 135.395 persons have been entered into this database[3].

J. Ehmer writes about the advantages of this type of source for research about artisans: The whole population is covered in population censuses and therefore restrictions because of the status within their occupation (apprentice, journeyman, master craftsman) can be overcome and information about their occupation can be combined with information about other characteristics. More than one census list would even allow to investigate stability and change (Ehmer 1987: 48). On the basis of such census lists he compared the cities of Zurich, Vienna and Zagreb. Master craftsmen had higher percentages of local born people than apprentices and journeymen. Better economic conditions lowered their share, while economic stagnation led to restrictions for migrants in becoming master craftsman. The highest percentages of locals are to be found among the higher status trades and trades with more financial needs. The migration history of master craftsmen mirrors an earlier stage in time than that of journeymen and apprentices and such facts have to be considered in comparing their places of origin (Ehmer 1987: 50-52).

1 A description and evaluation of this source can be found in: Nicholson 1999.
2 http://www-gewi.kfunigraz.ac.at/suedost/seiner/index.html
3 http://www-gewi.kfunigraz.ac.at/suedost/seiner/availability.html

The overwhelming majority of the apprentices was not born in Zurich or Vienna and only half of them in Zagreb. Most of them came from the surrounding region of Zurich and Zagreb, while apprentices in Vienna came predominantly from Bohemia and Moravia in mid 19th century (Ehmer 1987: 53f.). Journeymen born in the place of employment constituted about a fourth in the cities of Vienna and Zagreb, while in Zurich almost none of them was born there (Ehmer 1987: 54). Some trades had completely different regions, where they recruited their workforce, like bricklayers from Tyrol in Zurich or some textile artisans, which were mostly locals from Vienna (Ehmer 1987: 63-65).

Migration and GIS (geographic information system)

Migration is about movements of people in space and therefore a combination of the geographical information with the data of the Albanian population census of 1918 makes sense. Such a project could bring innovative aspects to the historical research of migration since historians are much concerned about time, but less concerned about space. The advantage of maps over the use of tables in the representation and research of spatial characteristics is that you can see spatial patterns more clearly. The researcher is not forced to aggregate data to larger spatial units, but can use individual data. You are not bound to majorities within characteristics, you can also investigate minorities and deviant characteristics. Thematic maps are not only a means of representation of data, but also a means of research. A written text has a linear structure, which is less well suited for the discussion of places, regions and geographical connections. Maps can, in contrast to written or spoken text, submit several statements and connections at the same time and combinations of them can lead to new results (Steidl 2003: 28).

O. Domonkos uses maps in his study about migratory movements of journeymen in Hungary in the 19th century, but in the maps there is no information about the number of people originating from different birth places (Domonkos 1987). A. Steidl uses 22 maps for showing the places of origin of artisans in Vienna, but also in her maps there is no information about the importance of the respective places of origins (Steidl 2003).

A rather new possibility, Historical GIS (geographic information system), can overcome these restrictions. A GIS can be defined as 'a sophisticated database management system designed for the acquisition, manipulation, visualization, management, and display of spatially referenced (or geographic) data.' (Aldenderfer 1996: 4). GIS can 'make spatial analysis easier, more flexible, and more powerful.' (Goodchild 1996: 241). This method of research depends on the quality of the geographical information it uses and the special character of geographic data: 'Geographical information is primarily cross-sectional – that is, it represents a 'snapshot' at a point in time. Mechanisms of cause and effect, on the other hand, are strongly temporal, and it is consequently difficult or impossible to deduce cause and effect from patterns observed in geographical data. Spatial analysis is primarily analysis of form, whereas understanding requires analysis of process.' (Goodchild 1996: 245). Geographic information is multidimensional, requiring at least two coordinates to define location, which distinguishes it from other kinds of information (Knowles 2002: xii).

In using historical data we are adding a new dimension to GIS and we combine geography and history, which can be distinguished according to the following statement: 'Geography is the study of spatial differentiation, history is the study of temporal differentiation. Historical GIS provides the tools to combine them to study patterns of change over space and time' (Knowles 2002: xii). Many sources used by historians contain geographic information, but much of it has not been used for research yet. Historical GIS projects are still very rare, some examples for it are the Great Britain Historical GIS project[4], the China Historical GIS[5], or the National Historical GIS for the USA[6]. For Southeast Europe there have no such projects been started yet. The more frequent use of GIS as a tool would allow scholars to explore the spatial aspects of history to a much higher degree. Migration studies are especially suited for such studies since migration is about movements of people in space and time.

The census data of 1918 is a rich source for a variety of questions related to studies about population structure and behaviour. Birth place and the present residence were registered for each person and therefore data for migration studies is available. The most

4 http://www.geog.port.ac.uk/gbhgis
5 http://www.people.fas.harvard.edu/~chgis/
6 http://www.nhgis.org/

important obstacle is the fact that there is almost no temporal information available about the migration of the people. Therefore we have to rely upon indirect information as comparisons of birthplaces of parents and children, changes of birthplaces among siblings, and comparisons of people of different age-groups (cohorts) about percentages living at their birthplaces. Since all the information is still on the level of individuals and not yet aggregated, there are much more possibilities for research compared with using published aggregated data. Data on the individual level allows us also to do research about the families of migrants and about chain-migration.

A first migration project

The research project 'Migration in Albania at the beginning of the 20th century'[7] in the years 2002 and 2003 was funded by the Austrian National Bank and based at the University of Graz, Department of Southeast European History. Data for the places in Albania was already existing due to the aforementioned project of data entry. The geographical information was derived from several maps of the first half of the 20th century because a part of the villages appears only on one map and the combination of several maps increases the number of digitised villages. F. Seiner, the director of the census, made some maps of the administrative division of the country (Seiner 1922: 13) and the territories of the tribes in the north of the country (Seiner 1922: 102), which have been digitised, too.

The project is using ArcView8 for building the Historical GIS. Places are digitised as points, which should be at the centre of the settlement. 1,557 settlements out of 1,794 have already been digitised, which are 86.8 percent of the settlements with 91.9 percent of the population. Recently acquired maps through a donation will hopefully increase the coverage of settlements to at least 95 percent in the course of the coming year.

Seiner's maps have already caused some troubles since they do not fit to the digitised geographical maps and to each other exactly. This is mainly because they contain no direct georeference. Other troubles are caused by ambiguous place names or illegible information. This project was designed as a migration project and therefore also places outside of the territory of the population census had to be digitised (place of birth of immigrants and place of residence of temporal or permanent emigrants).

Migration in Albania in the beginning of the 20th century followed several different patterns: There are differences between men and women, between cities and rural settlements. The percentage of rural adult men still registered in their place of birth is rather stable between 80 and 90 percent throughout adulthood. The pattern for married and widowed men is rather similar, while single men tend to be more mobile at higher ages. Only 40 percent of rural married women are registered in their place of birth. This is obviously the effect of patrilocal residence after marriage. The percentage of widowed women registered in their place of birth is almost 50 percent, which indicates the return of some widowed women into the household of their father or brother. This is especially pronounced for widowed women in their twenties. There is also a trend for unmarried women to leave their place of birth at higher ages.

In the cities different patterns for adult men can be observed: While a higher age normally meant a higher percentage of immigrants, the influence of the marital status was different. Overall the percentage of immigrants was lowest in Kavaja and Tirana with 10 to 20 percent, in Shkodra with 40 percent, and in Durrës up to 60 percent at higher ages. Unmarried urban women were more likely to be born in the city than widowed and especially married women. This pattern is similar to that of rural women, but the percentage of women registered in their birth-place was much higher: about 90 percent in Tirana, about 70 percent in Kavaja and Shkodra and 50 percent in Durrës.

Occupational migration to cities

There were only seven cities in Albania in 1918. The data for one city has been destroyed and there is obviously a problem with the data concerning the birthplace for two other cities. The data about urban migration is therefore available only for four cities: Shkodra (pop. 23,099), Tirana (pop. 10,251), Kavaja (pop. 5,453), and Durrës (pop. 4,175, Seiner 1922: 7). Immigration to Shkodra in Northern Albania was mostly from the villages to the south-east of the city and the cities of Podgorica and Ulcinj from Montenegro. Immigrants to the coastal city of Durrës were born mainly in the surrounding villages and the cities of Kavaja and Ti-

7 http://www-gewi.kfunigraz.ac.at/suedost/seiner/migration_project.html

rana. The birth places of the immigrants to the city of Kavaja were much more evenly distributed in the surrounding settlements. Immigration to the Central Albanian city of Tirana came from the surrounding villages and the cities of Durrës and Debar (Macedonia). Migration from the surrounding villages was mostly a short distance migration up to 10-15 kilometres, while inter-urban migration reached 50-60 kilometres.

Research about migration of different occupational groups is restricted to 25 occupations[8] (the most popular), which means that at least 73 men were engaged in this occupation. The occupation with the highest number of people engaged in it, is that of a trader: 1,978 men were registered under such occupational titles (16.0 percent of the adult male population[9]). These 25 occupations amount to 59.6 percent of the adult male population. An additional 16.2 percent of them are not coded yet and for 8.3 percent of them there is no information about their occupation. 74.1 percent of these men were born within the city, which is slightly lower than the share of the other adult men (77.3 percent). Kavaja and Tirana have rates of about 90 percent, Shkodra's rate is 67.1 percent and in Durrës it is lowest at 56 percent.

Immigration differed considerably between these occupations: The highest rates of adult men born within the city were among farmers (93.5 percent), ministers of religion (88.3 percent), shoe-makers (87.7 percent), barbers (87.5 percent), teachers (86.5 percent) and blacksmiths (86.3 percent). The highest share of migrants to the city are to be found in the following occupations: seaman (83.0 percent), bricklayer (81.0 percent) and soldier (52.6 percent).

There exist considerable differences within one occupation across different cities[10]: Half of the traders and owners of cafés, shoe-makers and barbers were migrants in Durrës, while in the other cities their proportion was less than a fourth. Two thirds of domestic servants, bakers and policemen were migrants in Durrës, while in Kavaja less than 10 percent were migrants. In Shkodra three fourth of the agricultural day-labourers were migrants, while in Tirana only 1.4 percent of them were migrants. In Shkodra and Durrës almost half of the butchers were migrants, while in Kavaja and Tirana less than a sixth of them were migrants. In Shkodra more than half of the labourers[11] were migrants, while in Kavaja less than 10 percent were migrants. In Shkodra more than half of the day-labourers were migrants, while in Tirana a mere 5.5 percent of them were migrants. Those men living off their property or pension were mostly migrants in Durrës, but in the other cities migrants were only 15 percent or less of them.

209 traders registered in Shkodra were born outside of the city: 88 of them were born in Podgorica and 48 were born in Ulcinj and the others came mostly from Northern Albania and Montenegro. They are therefore relatively similar to the overall pattern of migrants to Shkodra. Domestic servants and porters were coming from all over Northern Albania, while policemen were born in Northern Albania and in Kruja. Soldiers were coming from Northern Albania and also from outside Albania: The birth place with the highest number of soldiers born there was Debar in Macedonia. Agricultural day-labourers were coming mostly from Montenegro, other day-labourers were coming from Northern Albania and Montenegro. Very high concentrations of birth places were among butchers: 63 out of 79 migrants were born in Podgorica. A similar concentration was among seamen with 138 out of 156 coming from Ulcinj. Bricklayers were very prominent among migrant labourers and half of them coming to Shkodra were born in Debar in the zone of migrant labourers.

Immigrant traders and labourers to Durrës were similar to the overall migration to this city, as was the immigration of labourers to Kavaja. Only 21 traders in Tirana were born outside of this city, 12 of them in Debar. Domestic servants were coming mostly from the villages around Tirana.

Conclusion

Several groups of occupations can be differentiated according to their migrational background:

– migration similar to the overall migration: traders and labourers

8 Occupations are coded according to the HISCO-scheme (van Leeuwen, Maas, Miles 2002). Up to now only such occupational titles are coded, which have been registered at least ten times.

9 Adult male population means here men of at least 18 years of age.

10 Here are only occupations with at least ten people in the respective city considered.

11 Men, who were registered simply as 'labourers'.

– migration mostly from villages and short distances: domestic servants, porters, policemen, soldiers and day-labourers
– migration with larger distances and no high concentration: agricultural day-labourers
– high concentrations on one birth place: butchers, seamen, bricklayers

The study of occupational migration to the cities of Durrës, Kavaja, and Tirana is hindered by the relatively low number of cases in almost all of the occupations. Shkodra as the largest city in this area is therefore much better suited for such a study. Future research will have to concentrate on larger occupational groups to get statistically more relevant results.

References

Mark Aldenderfer (1996) Introduction. In: Mark Aldenderfer, Herbert D. G. Maschner, eds., Anthropology, Space, and Geographic Information Systems (Spatial Information Series), New York, Oxford, pp. 3-18.

Ottó Domonkos (1987) Zur Wanderung ungarischer Gesellen im 19. Jahrhundert. In: Klaus Roth, ed., Handwerk in Mittel- und Südosteuropa. Mobilität, Vermittlung und Wandel im Handwerk des 18. bis 20. Jahrhunderts (Südosteuropa-Studien, vol. 38), München, pp.69-85.

Valentina Duka (1997) Qytetet e Shqipërisë në vitet 1912-1924, Tirana.

Josef Ehmer (1987) Die Herkunft der Handwerker in überregionalen städtischen Zentren: Zürich, Wien und Zagreb zur Mitte des 19. Jahrhunderts. In: Klaus Roth, ed., Handwerk in Mittel- und Südosteuropa. Mobilität, Vermittlung und Wandel im Handwerk des 18. bis 20. Jahrhunderts (Südosteuropa-Studien, vol. 38), München, pp. 47-69.

Josef Ehmer (1988) Gesellenmigration und handwerkliche Produktionsweise. Überlegungen zum Beitrag von Helmut Bräuer. In: Gerhard Jaritz, Albert Müller, eds., Migration in der Feudalgesellschaft, Frankfurt/Main, pp. 232-239.

Michael F. Goodchild, Geographic Information Systems and Spatial Analysis in the Social Sciences. In: Mark Aldenderfer, Herbert D. G. Maschner, eds., Anthropology, Space, and Geographic Information Systems (Spatial Information Series), New York, Oxford, pp. 241-250.

Hermann Gross (1987) Die Entwicklung des Handwerks in Südosteuropa unter mitteleuropäischen und osmanischen Einflüssen. In: Klaus Roth, ed., Handwerk in Mittel- und Südosteuropa. Mobilität, Vermittlung und Wandel im Handwerk des 18. bis 20. Jahrhunderts (Südosteuropa-Studien, vol. 38), München, pp. 11-21.

Patricia Kelly Hall, Robert McCaa, Gunnar Thorvaldsen, eds. (2000) Handbook of International Historical Microdata for Population Research, Minneapolis.

Karl Kaser (1992) Hirten, Kämpfer, Stammeshelden. Ursprünge und Gegenwart des balkanischen Patriarchats, Wien.

Karl Kaser, Robert Pichler, Stefanie Schwandner, eds. (2001) Die weite Welt und das Dorf. Albanische Emigration am Ende des 20. Jahrhunderts (Zur Kunde Südosteuropas, Albanologische Studien, vol. 3).

Russell King, Nicola Mai, Stephanie Schwandner-Sievers, eds. (2005) The New Albanian Migration, Brighton, Portland.

Anne Kelly Knowles (2002) Introducing Historical GIS. In: Anne Kelly Knowles, ed., Past Time, Past Place: GIS for History, Redlands, pp. 1-18..

Sonila Kora (2000) Albania: The Albanian Emigration in U.S.A. In: Southeast European Joint History Project 'Migration in the Balkans: Conflict and Integration' Thessaloniki, Greece, 12th-13th May, 2000.

Michael Mitterauer (1985) Gesindedienst und Jugendphase im europäischen Vergleich. In: Geschichte und Gesellschaft, vol. 11, pp. 177-204.

Beryl Nicholson (1999) The census of the Austro-Hungarian occupied districts of Albania in spring 1918. A preliminary note on the manuscript (Centre for Scandinavian Studies Papers No. 5).

Osman Okyar (1987) Industrialisation and Handicrafts in the 19th Century Ottoman Empire. In: Klaus Roth, ed., Handwerk in Mittel- und Südosteuropa. Mobilität, Vermittlung und Wandel im Handwerk des 18. bis 20. Jahrhunderts (Südosteuropa-Studien vol. 38). München, pp. 183-193.

Michael Palairet (1987) The Migrant Workers of the Balkans and their Villages (18th Century –World War II). In: Klaus Roth, ed., Handwerk in Mittel- und Südosteuropa. Mobilität, Vermittlung und Wandel im Handwerk des 18. bis 20. Jahrhunderts (Südosteuropa-Studien vol. 38). München, pp. 23-46.

Roger S. Schofield (1987) Age-Specific Mobility in an

Eighteenth-Century Rural English Parish. In: Peter Clark, David Souden, eds., Migration and Society in Early Modern England, London, pp. 253-266.
Franz Seiner (1922) Ergebnisse der Volkszählung in Albanien in dem von den österr.-ungar. Truppen 1916-1918 besetzten Gebiete (Schriften der Balkankommission, Linguistische Abteilung, vol. XIII). Wien, Leipzig.
Spiro Shkurti, Karl Kaser (1997) Der Mythos vom Wandervolk der Albaner. Landwirtschaft in den albanischen Gebieten (13.-17. Jahrhundert) (Zur Kunde Südosteuropas III/1). Wien, Köln, Weimar.
Annemarie Steidl (2003) Auf nach Wien! Die Mobilität des mitteleuropäischen Handwerks im 18. und 19. Jahrhundert am Beispiel der Haupt- und Residenzstadt (Sozial- und wirtschaftshistorische Studien, Institut für Wirtschafts- und Sozialgeschichte, Universtität Wien, vol. 30), Wien, München.
Nikolai Todorov (1983) The Balkan City, 1400-1900 (Publication on Russia and Eastern Europe of the School of International Studies, University of Washington, vol. 12). Seattle, London.
Marco H. D. van Leeuwen, Ineke Maas, Andrew Miles (2002) HISCO Historical International Standard Classification of Occupations, Leuven.

Modern information retrieval technology for historical documents

*Markus Heller, M.A. & Dr. Georg Vogeler**

Scope

Historical information is fleeting: documents are subject to corrosion and their lifetime is not likely to prolong each time they are handed out to a researcher. In addition, originals can only be made available to one person at a time and editions may have most various formats. Nevertheless, technical development in the recent decade has improved the opportunities for historical research tremendously: Through the internet several researchers may access historical corpora at the same time, no matter where they have been compiled and made available. Due to the immaterial nature of electronic historical corpora the documents are not liable to quality loss –an advantage many organizations have recognized long ago, even though one has to admit that the long term storage is still an open issue.

Many projects, targeting the preservation and electronic encoding of documents have been funded, with a great degree of activity going on. Though, many of these projects must be stated to operate like independent islands, lacking common encoding standards and behaving as though the internet revolution had not taken place. Indeed, there is no attempt to integrate these individual corpora into a greater on-line available corpus of international historical heritage whatsoever. For this reason, the 'THING' project has been set up: THING (‚This Is Not Google')[1] attempts to develop a crawler, an indexer and a search frontend for online available corpora that make use of the XML markup as specified by the Schema of the Charters Encoding Initiative (CEI)[2]. As opposed to the capabilities of modern internet search engines, THING does not only record and index the documents in a flat manner. Instead, the corpora can be searched by XML tags: *e.g.*, a specific search for a location, mentioned in an editor's note, would only yield the set of documents where this condition is given.

Another important issue needs to be mentioned: Already now historians' most difficult problems reside in the issue of finding the appropriate document or the according class of documents, preferably sorted by relevance. The benefits of internet search engines are generally known, though no attempt has been made to transfer the ranking and sorting capabilities of these systems to the field of historical research.

Our project intends to develop and provide such a search engine, granting online access to ranked and sorted selections, taken from online available corpora of historical charters. We will thus provide the infrastructure and portal for future historical research in world-wide distributed electronic charters corpora.

CEI: The evolvement of a standard for historical documents encoding

Charters are central sources for the history of the period between the demise of the Roman Empire and the discovery of America. Before the middle of the 15th century legal documents were rare individual items. That is why charters, deeds, writs etc issued by emperors and kings, popes, princes, bishops, notaries and others, which have survived since the Middle Ages, are of such paramount importance as historical sources to all scholars.

Only very few such documents created before the 12th century have survived and these therefore often

*Ludwig-Maximilians-Universität München, Germany

1 Also known by its German name: DING (Das Ist Nicht Google), www.cis.uni-muenchen.de/~heller/Classes/ding/index.html

2 http://www.cei.lmu.de

are the only existing record of a particular historical event. With the beginning of the 12th century, it became a lot more common to lay down legal rights in written documents. It is therefore possible to do statistical evaluations of a great variety of daily legal transactions for research in the fields of legal, social, economic history or in order to gain a better insight into the mindset of the people living at the time. However, it is not only historians in their fields of research who work with these documents. Often the oldest records of European vernacular languages can be found in these charters, which makes them an outstanding source for linguistic research. All this means that medieval charters provide excellent material for many fields of research.

There have been many attempts to transfer these resources to the digital world. Since the 1990s scholars in the field of diplomatics have tried to build digital corpora. While older projects were based on proprietary software (*e.g.* the CD of the 'thesaurus diplomaticus' in Belgium[3], on HTML (*e.g.* Stuart Jenks and Jürgen Sarnowsky in their Preußisches Urkundenbuch[4], or on relational database management systems (*e.g.* the Anglo-Saxon Charters on the World-Wide-Web)[5], more recent projects such as the Codice Diplomatico della Lombardia Medievale[6] or the Cartuliare blanc of the Ecole des Chartes[7] use XML as their technological basis. In April 2004 scholars representing a good many of these new attempts using XML met in Munich to talk about the possibility of bringing these projects together. In order to consolidate the different approaches, they decided upon the foundation of a working group called 'CEI' (Charters Encoding Initiative)[8] and have agreed on defining a common semantic.

It has been observed[9] that historical sources are more than simply text: Traditional edition science recognizes them as testimony of historical events, but also as relics from the past and these properties have come to be described in more or less formal ways. A search engine would consequently have to consider these metadata in their according structure.

Since the middle of the 19th century diplomatics, the scholarly discipline dealing with medieval and early modern charters, has developed a well founded system for such metadata. Scholars have established principles for scientific editions as well as for calendars collecting the summaries of charters. With the beginning of the use of computers in the humanities theoretical approaches have been made to emphasize synergies[10] and to make use of the prior archievements from the print era[11]. Considering the continuously growing number of digitally encoded charters collections[12], Sahle and Vogeler have proposed the setup of a charters portal in the internet which would assemble the numerous encoding approaches into one common XML standard for the digital encoding of charters.[13]

As said, the CEI attempts to find semantic standards therefore, albeit one has to concede that reality is still far away from this goal. Nevertheless, due to the high usage of formal methods in charters research in the past century, very similarly structured standards have come into existence. They all follow similar annotation patterns, even though they apply tagging with a differing degree of detail: Whereas Annegret Fiebig's cor-

3 Demonty, Ph. ed. Thesaurus Diplomaticus, Turnhout: Brépols. 1997.
4 http://www.rrz.uni-hamburg.de/Landesforschung/orden.html
5 http://www.rrz.uni-hamburg.de/Landesforschung/orden.html
6 http://cdlm.unipv.it
7 http://elec.enc.sorbonne.fr/cartulaireblanc/
8 http://www.cei.lmu.de
9 Constantopoulos, P. et al. (2002): Historical documents as monuments and as sources, Computer Applications and Quantitative Methods in Archaeology Conference, CAA2002, 2-6 April, 2002, Heraklion, Greece. http://www.ics.forth.gr/isl/publications/paperlink/caa2002.pdf (accessed 27 November 2004).
10 Karsten Uhde: Urkunden im Internet –Neue Präsentationsformen alter Archivalien, in: AfD 45 (1999) , S. 441-464.
11 Michele Ansani: Diplomatica (e diplomatisti) nell'arena digitale, http://dobc.unipv.it/scrineum/ansani.htm, in: Scrineum 1 (1999), S. 1-11. = Michele Ansani: Diplomatica (e diplomatisti) nell'arena digitale, in: Archivio storico italiano 158 (2000), S. 349-379. See also: Patrick Sahle, Thorsten Schaßan: Das Hansische Urkundenbuch in der digitalen Welt. Vom Druckwerk zum offenen Quellenrepertorium, in: Hansische Geschichtsblätter 118 (2000) , S. 133-155.
12 http://www.vl-ghw.lmu.de/diplomatik.html
13 Patrick Sahle u. Georg Vogeler: Urkundenforschung und Urkundenedition im digitalen Zeitalter, in: Geschichte und Neue Medien. Kongreß .hist2003, Berlin April 2003, Berlin 2005 (Historisches Forum. Schriftenreihe von Clio-online 5), to be printed.

pus[14] annotates the original punctuation, the Regesta Imperii[15] focuses the summary with great detail.

A search engine that intends to cover all these different corpora will have to consider the according structures and try to match and convert them to a single standard, so that queries may be successful all over the heterogenity of the different corpora. For this reason we intend to use the CEI as a reference structure and render documents from the Codice Diplomatico della Lombardia Medievale (Universita Pavia), the Bavarian Books of Tradition (Bayerische Traditionsbücher, Ludwig-Maximilians-Universität, München), the Chartular of Corbie (Ecole Nationale des Chartes, Paris) and the Diplomatarium Norvegicum (University of Oslo) searchable through a portal.

Following the stated distinction between metadata and the very text, also the CEI methodology distinguishes two blocks of data for each document: Metadata and the actual text. The metadata hierarchy contains normalized information: a summary, the assignment of a date of origin, the description of the material form, the tradition of the text in modern and comprehensible language, including numbers and dates as used today. The actual text, the '<tenor>' of the document, contains the original with its internal structure which is given from the formalized judicial language as used thence: It follows the formal customs of the times by producing an introduction, naming reporters, issuers, receivers, a statement of affairs, means of authentification and certification. They thus adhere rigidly to a framework that allows to be annotated by using the CEI.

Modern search engine architecture and CEI corpora

Modern search engines are subject to two fundamental challenges: a) The sheer amount of information that needs to be processed and b) the impressively short response times. Making use of elaborated indexing and data preprocessing technology, they are able to provide the most perfect fit to the intended query at zero time. Search engine research however has gone beyond simply indexing the keywords as they appear in the documents: As the user may not specify his search interest as explicitly as given in the targeted document, the search engine must ensure that the document will be presented nevertheless. This development can be summarized by the term 'verticality' as opposed to horizontal, general internet search engines that are not specialized on any particular purpose: A vertical search engine considers the specific properties of the corpus and the requirements of the intended usage and deployment. For example, in order to meet the most specific demand, the search engine must entail a classifier for the according topics, combined with very specific tokenizing and lemmatizing capabilities.

In other words, it must possess surrounding background information in order to judge the target document as a 'god fit' to the query, even though the query keywords are not contained. Future historical search engines will face not just the need for verticality concerning the contents of documents. Instead, the challenge will be much higher as we deal with multilingual, diachronic corpora. Such a search engine will need to recognize that a document written in a medieval vernacular is also relevant to a query specified in modern German, English or Dutch.

There is another important distinctive issue that challenges historical search technology: Contemporary internet search engines will deliver documents that contain the query string just anywhere in their body. They don't provide the ability to select normalized dates, location and person names only, or charters and documents published by specific editors. We call this feature 'intra-node retrieval'. A historical search engine will have to be able to consider the XML layout as used by the publisher and match the structured query as given by the user to the structured set of documents in the index.

Yet, another issue is to be considered: Even though we generally encourage publishers to grant unrestricted read access to their corpora, some may require to impose access control and charge users for their research. In order to retain access, such a search engine will need to respect and inherit the access controls as defined by the corpus publisher. It will also need to inherit those controls to the engine users, demanding a quite distinct spectrum of access definition and restriction capabilities.

Before advancing to details, the overall architecture

14 Annegret Fiebig: Urkundentext. Computergestützte Auswertung deutschsprachiger Urkunden der Kuenringer auf Basis der eXtensible Markup Language (XML) , (Zugl.: Diss. Berlin 1996/97) , Leinfelden-Echterdingen 2000 (Schriften zur südwestdeutschen Landeskunde 33).
15 http://www.regesta-imperii.org

must be explained: A crawler component receives a list of so-called 'seed-urls' together with access information in case of access restrictions and downloads the html-pages recursively by following the links contained therein. In case it realizes that the downloaded file does not follow the HTML standard but contains XML code it tries to assess the according schema or DTD file. Being a module, it only tries to download files into a cache, which will later be fed into the index through a feeder. The cache, though, can also consist of locally available XML files, but be also a mixture of local and downloaded files.

The index is the core component of any search engine, no matter how the data storage questions are implemented. The documents will be processed long before the according query occurs and brought into a format that resembles the index of a book. Therefore important words and phrases ('tokens') need to be extracted from the text by very distinct methods and arranged in a way that query processing algorithms may find the according information on the documents of origin within a minimum stretch of time. The query terms must be subjected to the same methods as the tokens, in order to produce a good match. Answer aggregates will then be compiled and processed to provide information on the documents of origin. Yet, the answer set will have to be sorted according to the proximity between query and result. This method called 'ranking' will also have to rely on precomputed properties in order not to delay the delivery of the results.

Our search engine architecture also considers load sharing and clustering aspects. As we expect the number of available corpora and digital libraries to grow exponentially in the next years, we understood that modern search engine architecture must remain scalable in a linear way. For this reason we decided to introduce a superserver database in which all configuration options for the individual nodes is stored. A superserver on each node retrieves the configuration for each system and starts the individual components. It also controls the modules and triggers alarms in case of problems and keeps load information on all the search cluster modules in a central place.

Crawling aspects

A crawler serves as the interface to the corpora. It must be provided with a number of seed urls and information about the rough location of the published documents, together with information on access control. It will drill down into a hierarchy of web pages and identify pages of interest, here namely XML files that contain charters data. It will download the corpora into a cache and keep them prepared to be fed into the document processing queue. Since we expect not all charters corpora to follow the CEI definitions, certain other specific tagsets will still be convertible through XSLT since they target the same class of documents and also support their properties. This conversion will be supported by the system as there is an international vocabulary of diplomatic terminology that can be used as an ontology.[16] We expect to be able to convert certain incoming foreign but previously known XML standards to CEI, which is the overall standard in our index.

Our crawler will record access control information for every corpus, as corpus publishers will have a vested interest for their corpus to be made usable. In case they impost access controls on other types of usage, our system will respect these restrictions and the search results will only contain the link towards the data source. It will not offer the files directly for download from the crawler cache.

Indexing issues

Once the corpora have been retrieved, they need to be fed into the indexer. This is the place where the aforementioned document processing algorithms are used in order to allow intelligent matching during the query processing phase. After verifying the validity of the XML files against previously and manually installed schema or DTD files, the Feeder acts as a load balancer. Since there is a simple capacity limit for each indexer, it will consider the maximum load information from the central database and accordingly direct the documents to the node with the least load.

The receiver node will accept the document and forward it to a document processing black box. Even though the indexer module offers some very basic to-

16 Vocabulaire international de la diplomatique, ed. Maria Milagros Cárcel Ortí, 2nd ed., València 1997 (Collecció Oberta).

kenizing features, there is a way to override this setting and install very sophisticated mechanisms of artificial intelligence, very much depending on the type of the corpora. In our case, this module will act on properties of historical charters and support segmentation and tokenization as required by our intended user groups: Historians, Linguists and charters researchers.

Basically there are two ways of processing such data: On the one hand, it would be possible to reduce all items: During document processing, all terms could be stemmed and lemmatized, phonetic analysis could attempt to find basic forms and variations of person and place names could be standardized. This is a place for diachronic algorithms that would recognize 'Ratisbona' and 'Regensburg' as equivalent or person names such as 'eccehard' and 'Eckart' as related and substitute the original forms by normalized expressions. The other strategy would try to expand terms: It would enrich the node terms by assigning ontologically retrieved superclass information, such as that a specific term (person or place name) allows to refer the document to a specific epoch or to a certain political unit, or to a certain historical process.

These features are meant by the term 'verticality', in the sense that such a search engine would beat the best general internet search engines substantially if it comes to historical charters. On the other hand, such an approach would most probably not be competitive with engines like Google's in terms of scalability.

The strategies of reduction and enrichment are not exclusive: Instead, certain token properties such as flective endings would preferably be dropped, whereas significant terms, retrieved by algorithms such as TF-IDF[17] would be expanded to their superclasses, so that a match on the superclass would still produce the same document node. The degree of ion however would be a measure to determine the precision rank: A pure superclass match would yield a low match rank, a match on the word stem however would return a high relevance value.

Since the index needs to support structural queries in combination with pure content queries, a join of both query operations during query time would have been the traditional approach. We decided to make use of an algorithm called 'Content Aware DataGuide' (CADG), which precomputes structural and content access and thus obviates the expensive join. In tests on other XML corpora it has proven to exceed traditional methods of matching extraordinarily by speed, while keeping the file space requirements in manageable ranges.[18]

Query requirements

Whereas performance requirements on the indexing and document processing steps are moderate as they are not carried out at a moment where a user is actually waiting for a result, the query processing and index access is critical to the acceptance of a search engine. The query side can be split into two core fields of development: a) the index access and query processing and b) the user interface. The latter must render the usage of the search frontend as intuitive as possible, while not limiting the user in formulating his search conditions. As opposed to a flat index like Google's or Yahoo's an engine for semistructured corpora needs to offer an interface with which the user may specify structure constraints, next to the term entry. After all, it will have to guide the user through assembling a structure query following the XQuery[19] or XPath[20] or similar specifications that may be interpreted formally by the search module. The search module will have to receive the structural query string, process it, assign answer aggregates, subject them to a ranking procedure and deliver them in a format that can be converted into HTML.

As said, the user interface must not limit the scientific user in formulating his queries. We have identified three main groups of users for such a search engine: Historians, diachronic linguists and diplomatic researchers. They all may have different questions to the corpora. A historian may want to search for texts on a historical process or narrow down the query on a certain period of time, combined with a selection of a political entity. A linguist would want to select all docu-

17 Term Frequency and Inverse Document Frequency. For a survey on term weighting methods see: G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill Book \& Co., New York, 1983. See also: Ricardo Baeza-Yates, Berthier Ribeiro-Neto: Modern Information Retrieval. Harlow: Adison Wesley, 1999. 29 ff.

18 Felix Weigel, Holger Meuss, Klaus U. Schulz, Francois Bry. Content and structure in indexing and ranking XML. In: WebDB ,04: Proceedings of the 7th International Workshop on the Web and Databases. Paris, France. ACM Press, New York, 2004. 67-72.

19 http://www.w3.org/XML/Query

20 http://www.w3.org/TR/xpath

ments across a greater area, but very limited in the variability of the date of origin, in order to demonstrate the distribution of certain dialectological properties. The diplomatic researcher may want to select charters following very specific formalizations.

It has become clear that a search engine on historical documents will not only have to make use of information retrieval technology as proposed by computational linguistics, but also will have to support a very special users's perspective. Not only the data or the indexing algorithms define the exact layout of the architecture of such a search system, but also the intended usage, the very special interest of its users.

Future research

A search engine architecture as proposed is not merely a technical development: Instead it must be seen as an implementation of methods of diachronic corpus research. As an evaluation system it may serve as a data source for historical research, linguistics, geographical social research, lexicography, diplomatic research and possibly many other fields. As far as diplomatic research is concerned it has already become obvious that research on medieval and 16th century charters cannot be split into national phenomena. The spreading of the notary document combined with the phenomenon of the notary proof of authenticity can only be investigated with a European focus. The epoch of specialized diplomatics which focuses each document individually could be substituted by a new era of comparative diplomatics. In this sense we expect a broad spectrum of impulses in many fields and hope that the arrival of internet technology in the humanities will produce a boost that is similar to the advances in communication in everyday life.

References

Michele Ansani: Diplomatica (e diplomatisti) nell'arena digitale, http://dobc.unipv.it/scrineum/ansani.htm, in: Scrineum 1 (1999), S. 1-11. = Michele Ansani: Diplomatica (e diplomatisti) nell'arena digitale, in: Archivio storico italiano 158 (2000), S. 349-379.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto: Modern Information Retrieval. Adison Wesley, Harlow, 1999.

Constantopoulos, P. et al. (2002): Historical documents as monuments and as sources, Computer Applications and Quantitative Methods in Archaeology Conference, CAA2002, 2-6 April, 2002, Heraklion, Greece. http://www.ics.forth.gr/isl/publications/paperlink/caa2002.pdf (accessed 27 November 2004).

Demonty, Ph. ed. Thesaurus Diplomaticus, Turnhout: Brépols. 1997.

Annegret Fiebig: Urkundentext. Computergestützte Auswertung deutschsprachiger Urkunden der Kuenringer auf Basis der eXtensible Markup Language (XML), (Zugl.: Diss. Berlin 1996/97), DRW, Leinfelden-Echterdingen, 2000 (Schriften zur südwestdeutschen Landeskunde 33).

Patrick Sahle u. Georg Vogeler: Urkundenforschung und Urkundenedition im digitalen Zeitalter, in: Geschichte und Neue Medien. Kongreß .hist2003, Berlin April 2003, Berlin 2005 (Historisches Forum. Schriftenreihe von Clio-online 5), to be printed.

Patrick Sahle, Thorsten Schaßan: Das Hansische Urkundenbuch in der digitalen Welt. Vom Druckwerk zum offenen Quellenrepertorium, in: Hansische Geschichtsblätter 118 (2000) , S. 133-155.

G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill Book & Co., New York, 1983.

Karsten Uhde: Urkunden im Internet –Neue Präsentationsformen alter Archivalien, in: AfD 45 (1999), S. 441-464.

Vocabulaire international de la diplomatique, ed. Maria Milagros Cárcel Ortí, 2nd ed., Univ. de València, València, 1997 (Collecció Oberta).

Felix Weigel, Holger Meuss, Klaus U. Schulz, Francois Bry. Content and structure in indexing and ranking XML. In: WebDB ,04: Proceedings of the 7th International Workshop on the Web and Databases. Paris, France. ACM Press, New York, 2004. 67-72.

Internet addresses

http://cdlm.unipv.it/

http://elec.enc.sorbonne.fr/cartulaireblanc/

http://www.cei.lmu.de

http://www.cis.uni-muenchen.de/~heller/Classes/ding/index.html

http://www.regesta-imperii.org

http://www.rrz.uni-hamburg.de/Landesforschung/orden.html

http://www.trin.cam.ac.uk/chartwww/

http://www.vl-ghw.lmu.de/diplomatik.html

http://www.w3.org/TR/xpath

http://www.w3.org/XML/Query

Integrating structured and unstructured searching in historical sources

*Rik Hoekstra**

Introduction

In the past years, XML has become the advocated way to encode editions of sources in the humanities, following the TEI standard, and other more specialized standards derived from it[1]. In the discussion about standards, problem solving is usually focused on the encoding of the information within the sources. The accessibility of the encoded information is a matter of less concern; the implicit idea seems to be that once all information in an electronic version of a source, a document, a corpus or a collection is encoded properly, it will be accessible and there will be merely a question of presentation. After all, structure and presentation are separate in XML documents. Presentation may, of course, take many forms, but I would argue that making the encoded information accessible is a much more intricate problem, especially when we are dealing with a more complex source or collection of sources with several layers of information.

This paper focuses on how to make information in several layers available for users in an integrated form. I will take examples from our experience at ING (Institute of Netherlands History) and the solutions we have arrived at.

Multi layered sources

A relatively simple example may shed some more light on the question. In the publication of the 'Rijmkroniek Holland' (rhyming chronicle of Holland, one of the earliest and most important narrative sources about the history of Holland) we faced the challenge to publish the transcripts of different versions of the text so that users would be able to compare the differences in them. Not all versions were equally long, and in some cases only fragments had survived. In addition, there were images available of the different manuscripts and it was clear that the transcriptions of the text should be full-text searchable. It is not possible to present all views at the same time, this requires different 'modes' of access: one to compare arbitrary passages of the different versions, one to search and one to present the image of a page and its transcription. As the underlying material remained the same, switching modes is possible in the interface at all times. While we were preparing the publication, we realised these modes would not be the only possible ways researchers would want to use the manuscripts. For linguistic analysis, for example, online accessible documents are not very practicle because they do not allow for a statistical approach that requires analysis of the whole text or substantial parts of it. Therefore, we also made the transcripts of the manuscripts available for download in plain text and TEI-XML form.[2]

Preparing an edition like this cannot fully rely on XML tools but requires a variety of electronic tools including a web programming environment and a database. Of course, XML and XSLT were useful tools and clever use of Xpath and XSLT combined might have made much of the approach possible, but they would have slowed down the system considerably and the solution would have lacked flexibility because they are no real programming environments. In the approach we took, XML and XSLT were elements in what is getting

*Instituut voor Nederlandse Geschiedenis, The Hague, The Netherlands

1 *http://www.tei-c.org/*

2 *http://www.inghist.nl/Onderzoek/Projecten/Rijmkroniek*. It is relatively simple because the text has little internal structure. I do not present this case because of the original approach we took, as we tried to adopt the best practices of other full-text editions manuscript available, even if there is no standard way to publish electronic editions.

known as an XML-pipeline approach.[3]

To return to the central question of this paper, proper encoding of the material is essential to make it accessible, an additional analysis of the use-cases enabled by the different layers is also necessary. This should not so much cover all possible uses of the underlying material, because a proper edition is meant to enable researchers to do new research, but rather to reveal different levels, or modes, of access.[4] This question is especially important in the case of sources that are mainly made up of text. The main problem is that text will not reveal its content beyond what is literally contained in it. Or, to put it in another way, a text has no context of its own. It must be provided, either by a collection or by the editor who enriches the information.

In editing historical sources with a substantial textual content, there has always been a difference between the editing of the texts themselves and the indexes. Texts are effectively unstructured data sources while the indexes contain much more structured information. Indexes are not the only structured data belonging with texts: in a classic publication, for example, there may be tables of contents and information about the provenance of the texts and other information about them. Coming from the field of library science, these structured data are usually called metadata, irrespective of whether they say something about the texts or they contain information related to the texts. However, it is a bit cumbersome to lump metadata and extended information together, because this may obfuscate that they may serve a number of rather different purposes:

– they provide information about the textual material in the collection – these are metadata in a strict sense
– they make information that is contained in the texts separately accessible – this is the purpose of indexes
– they may also provide extended information to clarify parts of the texts – an index item, for example years of living and a function of a person

Often, editions contain different types of indexes, such as indexes for people, geographic names and keywords, but classic editions also may combine different index items into one index. Of course, both meta data and indexes make classic text editions much more accessible, but they lack searchability and flexibility.

Searching electronic editions

Many have claimed, and for a long time, that the electronic editing of sources would solve this problem, because 'everything' would be searchable at the speed of light. Now, searching text and searching data are very well possible. We can search through texts and perform statistical ranking of the results. To take an obvious example, 'googling' the web searching for a word or a combination of words, will yield many results. In this way we can locate information that before was unfindable or at least cost a lot of work and dedication to search and collect. The example is a bit superficial, as googling the web is not searching a research collection. However, it is easy to use a search engine to search through our own collections of texts and we will find words or string patterns, to use a more technical term, much more efficiently than we could ever in a book.[5]

This is also the case if we search indexes, especially if they are stored in an appropriate manner. For example, take people. We can search through a database and retrieve all data about people extremely fast. A database, moreover, may contain much more additional information about the entities described in it than a classic index did. It will still take a lot of time to compile all this information into a database, and that is an important consideration when setting up an electronic resource. Once information about entities is included, there is no way back and you will have to include that information for all entries in the database to create a consistent resource. But that is not our concern here.

3 For a definition and links see *http://en.wikipedia.org/wiki/XML_pipeline* – note however that in addition to the usual approach that focuses on XML standards centric pipelining, it may include many different tools.

4 On a side note it should be remarked that apart from considerations related to the content of the sources, this invariably also requires an economic weighting of possibilities and return of investment, as each decision to encode information will have consequences in terms of time and resources.

5 It is not uncommon to present XML as between structured and unstructured or semistructured. some more technical papers about the subject: Nilesh Dalvi, Dan Suciu, Indexing Heterogeneous Data September 27, 2003 <*http://www.cs.washington.edu/homes/suciu/tech-report.pdf*>; T. Lahiri, S. Abiteboul, J. Window, Ozone: Integrating Structured and SemiStructured Data, *International Journal of Digital Libraries*, Vol. 1, n. 1, April 1997, pp. 68-88; S. Raghavan and H. Garcia-Molina. Integrating diverse information management systems: A brief survey. IEEE Data Engineering Bulletin, 24(4):44-52, 2001.

Electronic editions have made searching a lot easier, but they did not solve all problems and even created some new ones. To understand what happened, we have to consider what information is stored and how, what information we search for and what information we would like to retrieve. Say, we want to find information about a person, for example ('Johan van Oldenbarnevelt'), or a city ('Amsterdam'), or a ship name ('Duifje'), or an institution ('States General'), or an official ('Ambassador of France'), but we search for string patterns either in our textbase or our more structured data. Searching is complicated, as we all know. If we take the 'King of France', for example, he may appear in many different ways in the texts we search through. Searching for 'king of france' in a corpus we will possibly also have to look for 'french king', '+french +king', 'french AND king' or 'king NEAR france', or for his first name, for example 'louis', and hope he will not occur in too many other constellations. This example is by no means exhaustive and it may require quite a number of different string queries to find all locations in our texts where the person 'Louis, King of France' is mentioned in our corpus. It depends on the nature of the corpus how many variations there are, the subject of the text – is it only about Louis or is he one of the actors in a much bigger play – but also on the purpose of the text, as legal texts are very different from literary texts for reading and for searching. I leave aside the question of ancient forms of language, or a compilation of texts from different times with different orthographies, or in a mixture of languages, not because they would be trivial problems, but because the complications they introduce are of a comparable kind. Perhaps these limitations may be overcome with sufficient ingenuity and perseverance on the part of the user of the resources, in the case of my institute usually researchers. However this requires quite a bit of prior knowledge about the contents and peculiarities of the data source, and it is always dangerous to take prior knowledge for granted.

A convenient solution to these problems is to use an electronic index to the material, if one is available. For the electronic edition of the 'Rijmkroniek Holland' mentioned above, we did just this. For the book edition of the text an index was prepared and we used it to provide better access to the online versions. If no index is available or if the corpus is prepared from scratch, it is a convenient solution to do so in XML because then text and index may be combined into one document. This might appear to be an ideal solution: we may query through one single information resource, using the appropriate tools. If, for example, we mark people, it is possible to generate dynamically a list of references to a person in a collection of documents. Because the source of the references is an XML element with an entity of its own, the link between index and source is always maintained. At least, that is what literature suggests. This solution is, indeed, convenient, but as usual, the challenge is much more complex.

Seventeenth-century resolutions and XML

I will elaborate the question using the example of the ING publication of the States General resolutions and the solutions we arrived at. The publication of the resolutions is a long standing research endeavour of ING. It is an edition of the decisions (resolutions) of the ruling body (States General) of the Dutch Republic in the sixteenth and seventeenth centuries. There are now 21 printed volumes, divided into three periods. This electronic edition is a sequel running for a period of five years, from 1626 to 1631, comprising some five thousand resolutions per year, which amounts to 25.000 for this project in total. The States General dealt with every subject that came before it, both internal affairs and foreign relations. It received people, petitions and letters from private people and institutions high and low, from within the country and from without. In its decisions (resolutions) the States followed the order in which the cases and questions were presented.[6]

The States General held meetings almost every day of the year with few exceptions, and minutes were taken for each of these meeting days (called *zittingsdagen* in Dutch). At the beginning of the meeting, the clerk noted those present and the provinces they represented, so each *zittingsdag* has a list of those present. Further, all resolutions were marked separately. Within the resolutions the editions marked all people, institutions, ship names, and geographical names mentioned. All these appeared in the original manuscript in different forms of writing even if the same person, ship, place or institution was meant, as was usual in the seventeenth century.

6 The public verions of the resolutions will become available from *http://www.inghist.nl,* but the application is now still under construction and not yet public.

The electronic edition had to reflect all these peculiarities, staying as close to the original sources as possible. It was relatively easy to devise a structure to encode the resolutions in XML. The '*staten generaal*' document type definition (DTD) defines a '*zittingsdag*' as its basic element. Each *zittingsdag* document holds a list of those present and all resolutions with all elements mentioned above marked. Following the practice of the latest paper editions, that ran to 1525, the resolutions are not transcriptions but summaries. However, the summaries stay close to the actual text and the names of people, ships and institutions are retained as they appear. We decided that entities like people and institutions in XML should be tagged using some form of identification. There were several editors working at the resolutions, in principle each was responsible for one year. We decided to store data about the people, the institutions and the positions held by people in a reference system. The editors could refer to the reference system and it provided identification for the entities in the XML documents. We did not use XML encoded information (documents or records) for this, but a relational database with a web interface, mainly because the data for the reference entities fitted the relational model much better and because it was easier for the editors to use a web enabled database.

In the XML zittingsdag documents we used an identification number assigned by the database for identification of institutes and people, but not for ship and geographic names. On the one hand, using a support system for people, institutions and functions was convenient, because it offered a central and independent system for information that was external to the resolutions, but important as a reference system. For example, the members of the important families in the Dutch Republic appear many times in the resolutions. Many of them had the same or very similar names and it is easy to mix them up. Identification numbers were unique and made it possible to keep them apart even if their names were equal. Moreover, many people were known by another name than their official family name. To take just one example, the person who appears in the resolutions as 'Noortwijck', was actually called 'Nicolaas van den Bouchorst', who was lord of Noordwijk, not an unusual practice in those days. On the other hand, using a support system also has some down sides, as combining editing the XML documents and inserting id-numbers from a database is not easy using an XML editor such as the XMetal commercial editor we used. While it is not clear why this should be the case, the reason for this seems to be that, following the XML standard itself, XML editors are document centric and not well suited for combining different resources.

Integrated searching

From the onset of the project, it was clear that for a publication of the edition we would need both types of information, the resolutions and the data from the support database, but now we had two data sources that were very different in nature. The main data source were the resolutions, mainly a text resource and as such unstructured information, but the support database held largely structured information. Even if the the database had been built up to support the resolutions, it also held information that was not available in the resolutions themselves, but could well be important in querying them. Integrating the two resources was not an option, for two reasons. Firstly, the two data models could not be mixed in an easy way. Secondly, they were two independent resources that held other information and that were conceptually different. An example may clarify this. The support database held information about people and their positions. The database made it possible to look up the other people who held the same position. In addition, the support database allowed for much more flexibility in retrieving a person because it held, among other things, different variants of names.

Taking all this into consideration, it was clear that the eventual application would have to integrate searching both information resources at the same time or separately, depending on the mode of viewing. We devised an interface that would accommodate several ways of browsing and searching the material:

1. browsing resolutions
2. full-text searching the text of the resolutions
3. searching for people, positions and institutions using the support database
4. combining searching for tagged elements and full text

Of course, only actual use of the application will show its full scope and limitations, but I will highlight some of its possibilities here, following modes outlined above.

Ad 1
Browsing the resolutions is rather straightforward and based on a calendar view. From an overview of the days of one of the years 1626-1631 users can click to a specific day.

Ad 2
Searching for text will locate words in the resolutions. Users get feedback about the words their search pattern will correspond to, each individual pattern is clickable to narrow the search. Note that many of the entities contained in the support database will be found in the resolutions when searching full-text. However, to repeat what I remarked above, full-text searching is much less precise and may return either too many or too few results, and even both at the same time as it will only locate exact string pattern matches. The result page from the full-text search will show a snippets from the resolutions containing the terms, and provides a quick overview of the various resolutions. Clicking on them presents a page with the resolution.

Ad 3
Searching for people, institutes and positions will return items from the database that correspond to to the search string. From the results page it is possible to get more information from the database or click to the locations in the resolutions. This will present a result page that is the same as that from a full-text search.

Ad 4
Complex searching enables the users to combine full-text searching with searches for people, institutions and positions, using a form. Users fill the search form by either entering their own search terms or picking from lists, that are generated from the support database. In addition, there are lists of geographic and ship names. The lists are generated from the electronic indexes of the *zittingsdag* documents that were generated when the documents were added to the application. Thus, they are a bit different from the items in the support database, as they have no ID numbers and are more like conventual book index items, comparable to book indexes. While they could also be found by full-text searching, the availability of lists provide users with pointers to terms that are in the index.

In the presentation of the resolutions, all tagged items in the documents are presented as hyperlinks. Clicking on them will direct users to a page with links to more information about them; the type of information it depends on the item. In the case of a person this is a page with information about him, including a link to the other resolutions where he is mentioned and the positions he is known to have fulfilled. A special case are the positions. There are no direct references in the resolutions to positions as such: all references are to individuals. If one clicks on the positions on the information page about a person, the system will present all other individuals who held this position and they may be traced in the resolutions. In this way, it is possible, for example, to retrieve all resolutions dealing with the ambassador of France or, if searching in positions, even all ambassadors.

This overview of the possibilities is by no means complete, but it was only my intention to give some insight in the possibilities of an information system that integrates searching structured and unstructured information. There are two points I would like to add here. First, it has to be remarked that using this approach results in an open system that may be extended in the future. In the application of the resolutions, for example, information about positions, institutions and people could easily be added or linked. At the moment ING uses the same data structures we employed in the support database in other databases, compiled for other research projects, the results of which will become available within a few years. Because there is a significant overlap in content, linking this information is very well possible, even if this can never been carried out automatically and human supervision is always necessary.

Second, the resolutions system as described has no subject access. This is a decision concerning the content mainly because devising a keyword system and assigning them would have been very time consuming and because, in the experience of ING, it is hard to maintain consistency in a keyword system and arbitrariness out of it. The idea was that the combination of summarized resolutions and full-text retrieval would provide an acceptable alternative.

Conclusion
Editions of historical documents often have several layers. The layers differ from source to source and collection to collection and making them accessible requires an different solution. In this paper I have mainly paid attention to integrating unstructured and structural information, or textual information and index in-

formation about elements in the text but external to the sources themselves. Making them accessible in an integrated system gives the electronic resource an added value that would never be possible in a paper resource, because of the flexibility of the electronic resources. In the example application of the resolutions of the seventeenth-century Dutch Republic States General we combined different levels of access and different modes of viewing into one open system. By way of conclusion, I would like to argue that combining different modes of viewing and access will take electronic publishing of historical sources to a higher level that will make electronic editions both very different from paper editions and more interesting and full-featured than most current electronic editions. To prepare editions like that, the systems designers do not only need an insight into the information content of sources, but they also have to analyse the possible ways users will access the information in the resource.

Editing and exploratory analysis of medieval documents by means of XML technologies

Aleksandrs Ivanovs & Aleksei Varfolomeyev***

The case of the documentary source complex Moscowitica-Ruthenica[1]

Preliminary remarks

Application of computer technologies in archaeography, historical source studies, and auxiliary historical disciplines can and should be conducive to editing and exploratory analysis of vast complexes of medieval historical records, as well as to synthetic operations – using historical evidences for reconstruction of historical facts. In order to achieve the aims mentioned above, there is a need for adequate research tools and techniques that can be applied to unique documentary sources, which lack clear, normalized structure (original textual divisions) and, sometimes, even discrete identifiable blocks of a text. At the same time, in the case of editing and studying a *complex* of documentary sources, the prerequisite for efficacious use of these tools is the creation of a full-text database, which includes all the documents that constitute the complex[2]. The design of a database should precisely reproduce the actual structure of the complex as well as the interconnections between the documents that form a definite system. On the one hand, such a database can be considered an efficacious instrument for searching, analyzing, and aggregating historical information of documentary records, but on the other hand, the electronic database can also serve as a model for a paper publication of historical records (Ivanov and Varfolomeev 2004). It is worth stressing that both the paper publications of documents and the historical record databases (electronic historical editions) promote and stimulate the circulation and use of documentary sources in historical research.

Common practices of editing and analysis of complexes of historical records (Ivanov 2004[a], 2004[b]) show that there are wide gaps between different exploratory and technological operations, which, as a matter of fact, should be performed systematically as a sequence of steps: creation of electronic versions of historical documents (data capture), representation and aggregation of documentary information, creation of adequate full-text databases and electronic historical editions, elaboration of research tools, search interfaces, and technologies for historians – the users of databases, historical data processing, preparation of paper editions. It seems that modern computer technologies give us an opportunity to perform all these operations simultaneously within the framework of one technological system based on a common approach to historical documents.

This paper presents the results of the case study of the applicability of XML technologies to editing (Ivanov and Varfolomeev 2004) and analyzing a medieval documentary complex 'Moscowitica–Ruthenica' (hereafter – MR). In the focus of attention is the problem of markup of document texts, which is based upon TEI (*Text Encoding Initiative*) encoding scheme, as well as structural analysis of documents by means of XSLT queries. Approbation of computer technologies in the case of the complex MR will provide historians and editors with the opportunity to apply these tools to analyzing and editing other collections of medieval records.

*Daugavpils University, Latvia
**Petrozavodsk State University, Russia

Documentary complex 'Moscowitica–Ruthenica'

Among the most valuable historical records of the Latvian State Historical Archives in Riga, there is a collection of documents, which provides historians with firsthand information about relations of Muscovite, Russian and Byelorussian lands and towns – Smolensk, Novgorod, Pskov, Polotsk, Vitebsk, etc., as well as Lithuania (later – Poland-Lithuania) with Riga, Livonia, Hanseatic League, and some German towns in the late 12th – early 18th centuries. The historical name of this documentary complex is MR. The documents that were initially kept in the so-called External Archives of the Magistrate of Riga constitute the kernel of this complex (in total – 787 documents)[3]. Owing to archival practices and historical traditions of document storage in the former Archives of the Magistrate of Riga, some of the documents of this complex were kept in the Internal (Secret) Archives of the Magistrate (about 80 documents)[4], as well as in the former Landrate Board of Lifland[5]. Although this documentary collection as a department of the Latvian State Historical Archives does not exist any more, its documents constitute the 'natural', historical complex of documentary records, which should be published and studied as a whole (Ivanovs 2004[c]).

The document nomenclature and the composition of the complex MR reflect the functions of the Magistrate of Riga[6]: after Riga had joined Hanseatic League in 1282, the primary sphere of activities of the Magistrate was the maintenance of close political and economic contacts with Russian, Byelorussian, and Lithuanian lands and towns, as well as mediation in the wholesale trade between these lands and Hanseatic League. As a result of documenting these functions of the Magistrate of Riga, the following kinds of historical documents and records emerged: international political and trade (commercial) treaties between Riga, German towns, on the one hand, and Russian, Byelorussian, and Lithuanian lands and towns, on the other hand; confirmations of bilateral treaties; instruments of ratification; letters patents; judicial documents; private business documents, etc. Most of the above-mentioned documents (*Urkunden*) are the objects of an auxiliary historical discipline – diplomatics (Kashtanov 1988, pp. 11–27; Goetz 2000, pp. 153–173). Besides these documents, there can be mentioned such types of historical records as official and private correspondence; letters of credence; letters of indemnity; letters of administration; passports; memoranda, instructions, orders, and resolutions issued by the Magistrate of Riga; charters of immunity; ambassadors' and envoys' reports and messages from Rus and Poland-Lithuania to the Magistrate of Riga and German towns; trade statutes and regulations of German offices (*Hof*) in Novgorod and Pskov; official letters written by Russian, Byelorussian, and Lithuanian officials, governors, voevodes, and bishops; records of proceedings conducted by the Magistrate of Riga; complaints and petitions presented to the Magistrate by Russian and Livonian merchants; different inventories, etc. It is evident that the variety of types and kinds of documentary records is considerable: the complex MR embraces almost all kinds of historical documents, which were typical of record-keeping, legal practices, and documentation of international relations both in German and Russian spheres of influence in the Middle and Early Modern Ages.

Owing to diversity of the documents as well as to differences in record-keeping practices, it is necessary to take a flexible approach to editing and exploratory analysis of these documentary records. It should be also taken into consideration that the documents of the complex MR were drawn up in different languages – Old Russian, Old Byelorussian, Old Polish, Latin, German (*Niederdeutsch, Mitteldeutsch, Hochdeutsch*), and Swedish. Paleographic and diplomatic features and peculiarities of the documents are also different. Editors usually took into account the above-mentioned peculiarities and language diversity of the documents; as a result, different archaeographic techniques were applied to editing documentary records of Russian and German origin[7]. Furthermore, documentary editions usually reflect priorities, which are inherent to a certain historiographic school and national tradition in historical writing. Such priorities make historians search for documentary illustrations (examples), which can be conducive to substantiation of the concepts accepted in their national historiography (Ivanov 2004[b], no. 4, p. 106). Thus, the documentary collection MR has never been comprehensively represented in the printed documentary editions, and the complex approach to editing and analyzing these documents has not been realized. Therefore a new complete edition, both a paper and an electronic one, seems to be urgent; at the same time, the creation of an electronic full-text database will be a real 'embodiment' of the complex approach, which has been declared in archaeography and historical source studies.

Computer-based technologies in source studies

In order to evaluate the potentialities of computer-based technologies in editing and studying the complex MR, general tasks of such a research project should be set. To the authors' mind, the tasks that can be carried out by means of computer-based technologies are the following:

- Studying the processes of emergence and development (evolution) of different kinds of historical records, which form the complex MR;
- Hypothetical re-creation of the system of record-keeping and document circulation in order to carry out the tasks of external and internal criticism;
- 'Dynamic' reconstruction of historical reality (facts), which exists in the evidences of historical records;
- Probabilistic reconstruction of the 'lost' historical records and facts, which, possibly, could be reflected in these records;
- Revealing the 'living reality' in formalized, standardized descriptions;
- Preparation of a printed and electronic documentary edition of the complex MR.

All the above-mentioned tasks can be carried out within the framework of exploratory operations, which should be performed in a consecutive order.

In the *first stage*, the text of every individual documentary record taken separately is being marked up[8]. The markup should be a multi-level one in order to combine quite different parameters and elements of the document body within one and the same markup scheme. Such a markup offers a historian a number of opportunities to search for historical information, while implementing arbitrary queries, and to perform different exploratory (analytical) operations, while carrying out the tasks of external and internal criticism (Shafer 1969, pp. 99–142).

In the process of marking up texts of documentary records, formalization and standardization of source information occur. However, original texts of documentary sources are not transformed, because formalized, unified, and standardized elements are used in order to mark up (designate) structural divisions and parts of the document body. Thus, these elements, which form the 'superstructure' of a text, can be considered an interpretation of a documentary record.

In the *second stage*, marked up documentary records are included into a complex (aggregated) database, and an adequate set of tools, which are appropriate for searching, aggregating, and many-dimensional analysis of historical information, as well as for cross-comparison of different documentary records of one and the same complex, is developed. Linkage of sources within a database provides historians with unique opportunities to verify evidences of individual sources and, as a result, to determine the credibility of evidences. Finally, the aggregated database provides a researcher with instruments for historical reconstruction and other synthetic operations (Shafer 1969, pp. 143–171).

The *third stage* can include some particular exploratory operations in order to carry out the tasks of a certain editing and research project. In the case of the complex MR, one of the most important tasks is the analysis and electronic representation of the structure (*Formular*) of the documentary records, which constitute the database. Though the internal structure of these documents has been in the focus of attention of diplomatists for a long time[9], examination of component parts of documentary records is still being conducted by hand, according to the 'classical' patterns of the 19th century practices. It is doubtful that any representative data can be gathered if the analysis of vast complexes of documentary records is made by hand. Thus, application of computer-based technologies in order to make such an analysis seems to be urgent.

To the authors' mind, the most appropriate computer-based technologies that can contribute to realization of a complex editing and research project are XML technologies, the advantages of which should be discussed separately.

XML technologies and TEI encoding scheme

XML (*eXtensible Markup Language*) is a set of rules for creation of new markup languages, which represent any text as a 'tree' formed of nested logical components. Such a tree, side by side with a table, displays a fundamental structure of data. It is evident that for representation of data of a certain kind (for example, descriptions of similar objects, which have identical attributes) the tables are more suitable, while in other cases (for example, representation of a family structure) the tree presents the data more logically and conveniently for the users. Thus, XML-documents, in addition to relational tables, offer new opportunities for representation of collections of structured data, and XML technologies, on the whole, become a new paradigm in database design.

An essential feature of XML-methodology is the use of standardized markup schemes. At the beginning stage of application of XML technologies, any academic community makes an effort to develop a certain encoding standard, which conforms to the specificity of initial data. In the Humanities, the most popular standard, which is widely used in projects of creation of electronic full-text collections, is TEI. Originally, this standard was meant for description of printed editions of literary works. The description required the division of texts into chapters and paragraphs, verses – into stanzas, as well as marking ends of pages, direct speech, etc. However, potentialities inherent in TEI standard are much wider: this standard can be used not only for creation of adequate electronic versions of printed editions, but also for identification and allocation of a system of logical fragments in different texts, as well as for linkage of these fragments. Thus, TEI can be considered an appropriate tool for analyzing and interpretation of historical records.

A long list of projects using TEI standard can be found on the site of TEI Consortium. There are also projects, the purpose of which is the creation of electronic collections as well as the development of the standards of description of medieval historical records[10]. As a rule, such projects offer very specific markup schemes based on TEI (Hansen 2002; Poupeau 2004). These schemes are adapted to specific features of certain texts and to certain tasks carried out by publishers and researchers.

MR documents encoding principles

For a long time in source studies and auxiliary historical disciplines, rather firm rules have been applied to description of medieval historical documents, to representation of their information, as well as to internal and external criticism and synthesis (Goetz 2000; Kashtanov 1988; Shafer 1969, etc.). Encoding historical records, these rules should be strictly observed. The multi-level structure of a documentary record should also be taken into consideration. Accordingly, at the upper level of encoding, 'meta-information' about a record – heading and date (with substantiation), document kind, language, storage place, information about previous publications of a document, some commentaries, etc. – is provided. The second (palaeographical) level embraces external attributes and features of a document: handwriting and its peculiarities, material on which document is written, defects, seals, watermarks, archival marks, postscripts, docketing, endorsements, etc. Information about the division of a manuscript into sheets, as well as information about the layout of the text can also been encoded within this level. At the third (diplomatic) level, general structural components of a documentary record should be encoded and represented within the markup scheme: introduction (*Protokoll*), the main text (*Kontext*), and concluding formula (*Eschatokoll*), as well as their subdivisions, such as 'invocation', 'intitulation', 'inscription', 'salutation', 'arenga', etc. (Goetz 2000, pp. 162–168; Kashtanov 1988, pp. 169–172). At the next markup level, smaller parts of a text – the clauses – are distinguished. The levels mentioned above can be supplemented with a few other levels, which represent stylistic, syntactic, and lexical characteristics and elements of a document. Accordingly, linguistic and office formulae (stable expressions, clichés, stock phrases, etc.), etiquette elements and phrases, individual turns of speech, descriptions (individual parameters, non-standard characteristics, and evaluations of facts and situations), as well as persons' names, place-names, dates, references to other documents, etc. are singled out. Finally, symbols and letters, which constitute the text, are encoded.

The encoding scheme, which is focused both on representation of the text and on different analytical and synthetic operations, gives us an opportunity to distinguish different objects at every level of encoding. Since the objects of lower levels are 'nested' (put) into the objects of higher levels, linkage between all these levels within the framework of one markup scheme seems to be evident. Such a hierarchical structure of a document is an optimum prerequisite for application of XML technologies.

The marked up document, the encoding scheme of which represents all the levels mentioned above, should be considered a primary document of a database. By means of XSLT and XSL-FO transformations, the primary document can be transformed into:

– HTML, PDF, RTF-files; these files allow to display the document on the screen of a monitor and to prepare printed editions of documentary collections;
– XML-files with one encoding level; such files provide a basis for analysis, synthesis, and interpretation of records. These exploratory operations should embrace both manual and automatic distribution and inclusion of elements of a text into various clusters determined by a researcher, as well as calculation of

frequency of appearance of certain attributes and elements of a text, revealing interconnections between fragments of different documents, etc. The above-mentioned exploratory operations can also be represented by means of TEI markup scheme in order to make such research techniques, as well as scientific data, available to academic community.

Some remarks on the problem of representation of symbols and diacritic marks, which are characteristic of medieval manuscripts, should be made. To solve this problem, methods elaborated for the project Menota[11] can be used. Accordingly, all the symbols that can be found in the complex MR should be represented in the primary documents of the database as XML-entities (&name;), which are linked to certain codes of Unicode standard. Diacritic marks should be considered as separate symbols. In this case, representation of diacritic marks and other symbols will correspond to Unicode. In order to display texts in different languages on the screen of a monitor simultaneously, a special font, which contains all symbols used in the project, should be created.

Structural analysis of Polotsk charters: A case study

The structural (*Formular*) analysis and, simultaneously, electronic representation of the Polotsk charters[12] by means of XML technologies reveal the advantages of XML encoding scheme and potentialities of these technologies. At first sight, the text of any charter consists of three parts: quite a formal introduction, rather a vast main part, and a very short concluding formula. If one tries to reveal the logical sequence of semantic fragments of the main part of a text manually, one and the same markup operation will be performed over and over again (for example, marking up the Xerox copies of documents). Instead of this laborious routine, XML-documents, which represent the texts of the charters, can be used. By means of a specialized XML-editor the structural fragments of the texts are tagged, using the `<div>` element with the attribute '`type`', which can acquire different values. The list of such values is created in the process of encoding. Then, by means of an appropriate XSLT query, the marked up (and encoded) texts are displayed on the screen of a monitor, and every certain tag `<div>` is marked with a certain color in order to compare the structural fragments of the texts visually. Other XSLT queries provide us with statistical characteristics and data: information about similar structural elements in different texts can be singled out and aggregated; the typical disposition of such elements within the document bodies can be modeled, etc. Finally, this case study presents the aggregated (probabilistic) internal structure of a Polotsk charter.

It should be also stressed that, using XML technologies (as well as any other database technology), conclusions and theses can be comprehensively verified, since they are based on the quantitative 'measurements' of their reliability.

References

Goetz, H.-W. (2000) *Proseminar Geschichte: Mittelalter. 2. Auflage*. Stuttgart.

Hansen, A. M. (2002) 'Text Encoding of Manuscripts: Danish Prayer Books from the 16th Century.' In: *Le Médiéviste et l'ordinateur. no.41. [On-line]* http://lemo.irht.cnrs.fr/41/m041-09.htm

Ivanovs, A. (2004[a]) 'Kompleksa 'Moscowitica–Ruthenica' ieviesana zinatnes aprite: arheografisks apskats [Introduction of the Complex 'Moscowitica–Ruthenica' into Scientific Circulation: An Archaeographic Review].' *Latvijas Arhivi*. Riga, no. 2: 47–85.

Ivanov, A. (2004[b]) ' 'Moscowitica–Ruthenica' v Latviiskom gosudarstvennom istoricheskom arkhive: istoriia formirovaniia kompleksa, sostav i vvedenie v nauchnyi oborot ['Moscowitica–Ruthenica' in the Latvian State Historical Archives: Forming of Document Collection, its Composition, and Introduction into Scientific Circulation].' *Drevniaia Rus'. Voprosy medievistiki*. Moscow, no. 3: 47–54; no. 4: 94-106.

Ivanovs, A. (2004[c]) 'Vestures avotu kompleksa rekonstrukcijas problema: kolekcija 'Moscowitica–Ruthenica' [Complex of Historical Sources: Problems of Reconstruction].' In: *Vesture: Avoti un cilveki. VIII*. Daugavpils. Pp. 49–57.

Ivanov, A. and A. Varfolomeev (2004) 'Ispol'zovanie tekhnologii XML dlia vvedeniia v nauchnyi oborot kompleksa dokumentov 'Moscowitica–Ruthenica' [XML Technologies in Introduction into Scientific Circulation of the Document Collection 'Moscowitica–Ruthenica'].' In: *Digital Libraries: Advanced Methods and Technologies, Digital Collection. RCDL'2004*. Pushchino. Pp. 285–289.

Kashtanov, S. (1988) *Russkaia diplomatika* [Russian Diplomatics]. Moscow.

Khoroshkevich, A., comp. (1977–1985) *Polotskie gramoty XIII – nachala XVI vv.* [Polotsk Charters]. I – V. Moscow.

Lappo-Danilevskii, A. (1920) *Ocherk russkoi diplomatiki chastnykh aktov* [An Essay on the Diplomatic Analysis of Russian Private Documents]. Petrograd.

LVVA – Latvian State Historical Archives (Latvijas Valsts vestures arhivs).

Poupeau, G. (2004) 'Réflections sur l'utilisation de la TEI pour coder les sources diplomatiques à partir de l'exemple du Cartulaire blanc de l'abbaye de Saint-Denis.' In: *Le Médiéviste et l'ordinateur. no.43. [On-line]* http://lemo.irht.cnrs.fr/43/m043-12.htm

Shafer, R.J., ed. (1969) *A Guide to Historical Method*. Homewood (Ill.).

Notes

1 The paper is written under a grant from the State Foundation of Cultural Capital of Latvia. The project of editing the documentary source complex 'Moscowitica–Ruthenica' has been supported by the Latvian State Historical Archives.

2 It can be stated that historical sources should be studied within a context of their 'natural' complexes. These complexes come into existence spontaneously: historical sources form definite systems in the course of performing their initial functions. The main features of the natural complex of historical sources are the following: common origin, close historical interconnections between them, hierarchical disposition within the system, etc. See Ivanovs 2004c.

3 See *LVVA*, fond 673, inventory 4, boxes (Kasten) 18–20.

4 *LVVA*, f. 8, inv. 3, capsula A, no. 14–18, 41, 72; capsula B, no. 42; capsula C, no. 1–11, 23, 27, 34, 43, etc. See also: f. 8, inv. 4, no. 6–58.

5 *LVVA*, f. 214, inv. 6, files 114–116.

6 See old inventories compiled in the 16th – 17th centuries: *LVVA*, f. 8, inv. 1, file 59 ('*Register der Privilegien der Stadt Riga auf Befehl des Rahts im J. 1507 angefertigt*'); *LVVA*, f. 673, inv. 4, Kasten 19, no. 258 ('*Verzeichnis vom Jahr 1599 über die im Rigaschen Archiv vorhandenen, auf den russische Verkehr und den Nowgoroder Hof bezüglichen aelteren Dokumente*'), etc.

7 Sometimes in one and the same documentary publication, editing technique is chosen depending on the language of the document. See Khoroshkevich 1977 – 1985.

8 In this editing project, the document markup scheme conforms to the division of document texts accepted in diplomatics. See Kashtanov 1988, pp. 170–174; Goetz 2000, pp. 162–169.

9 In Russian diplomatics, the analysis of internal structure of documents had become extremely sophisticated by the beginning of the 20th century. See Lappo-Danilevskii 1920, pp. 135–157.

10 For example, see CELT Project. http://www.ucc.ie/celt; Repertorium of Old Bulgarian Literature and Letters. http://clover.slavic.pitt.edu/~repertorium; MASTER Project. http://www.cta.dmu.ac.uk/projects/master, etc.

11 Menota (Medieval Nordic Text Archive). http://www.menota.org .

12 In this case study, the Polotsk charters drawn up in the second half of the 15th century are being encoded and analysed: *LVVA*, f. 673, inv. 4, Kasten 19, no. 2 (1474), 5 (1471), 7 (1471), 8 (1471), 10 (1464/65), etc. The selected documents are of the same kind and were written in the same period of time. Thus, these documents are comparable, since they form a sub-complex within the framework of the complex MR.

Temporal language models for the disclosure of historical text

*Franciska de Jong, Henning Rode & Djoerd Hiemstra**

1 Introduction

Historical and heritage collections consist for a considerable part of text and may incorporate diverse text types such as journals, archival documents, and catalogue descriptions. Because of the historical distance, access to this content is not straightforward. Historical variants of text are often more complex to identify and retrieve than modern variants. This is due to the less standardized spelling, the effect of ongoing language change and differentword (de)compounding principles. Moreover, more words are ambiguous because one or more meaning shifts may have occurred. Common fulltext search tools can only be applied successfully by users who are able to formulate queries with (a) knowledge of historical language and (b) insight in the relevant time spam from which the words have evolved. This paper explores techniques which may compensate for these linguistic obstacles: linking of contemporary search terms to their historical equivalents and 'dating' of texts.

We envisage to restore the diachronic relationship between terms which may be obscured by language evolution and usage, by applying statistical language models. These models may support the automatic detection of semantic similarities between words and word ambiguities, and they also allow to classify a text according to the time span from which it originates. This approach involves building temporal profiles of words as longitudinal sections in a reference corpus and temporal language models as cross sections.

In section 2 some detailed examples will be presented of the added value of this approach both for the accessibility of historical content and the detection of language change in relatively recent corpora from the news domain. In section 3 an overview of related work will be given, plus some technical background on statistical language models. Section 4 describes the proposed methodology in more detail, and some experiments for it in the news domain will described in section 5.

2 What temporal models can do?

Two types of language change will be distinguished here: evolution that is exemplified by a series of etymologically cognate word forms, and evolution exemplified by a single word form with a series of differentbut distinguished meanings. An example of the first type is the series of Dutch forms for the concept *PILGRIM: pelegrime, pelgrime, peelgrime, peilgrime, pilgrime, ..., pilgrijm, pellegrijm, peregrijm*. An example of the latter would be the semantic evolution for Dutch words such as *wijf* (English: woman; shifted connotations) and th*kabinet* (English: cabinet). For the sake of simplicity we will ignore here combinations of these types of evolution.

Users of a search system will typically know one or more contemporary forms associated with the concept they want to search for. They would be helped if the search interface was enhanced with knowledge about diachronically related forms that can be considered synonyms. This knowledge could be available via lexicographic resources, but could also be detected (semi) automatically by using information on collocations (words that tend to occur in the same syntactic phrase; fixed meaning; *e.g.*, *commit a crime; raise a question* and significant cooccurrence figures (for words that tend to occur in the same documents).

An early form for the concept PILGRIM is as likely to cooccur with (a form for) the concepts PILGRIMAGE or TRAVEL as a contemporary form. Language models can catch such dependencies per period, but via corpusbased normalization the temporal evolution of a concept profile could also be captured. Conversely, shifts in meaning will also bring shifts in cooccurrence

*University of Twente, The Netherlands {f.m.g.dejong, h.rode, d.hiemstra}@utwente.nl

figures. In principle statistical language model should therefore be able to help detecting diachronic synonymy.

A lot of research, including experimental work will be needed to deliver a proofofconcept for this idea, and to get a better grip on the question which information sources can exploited and or combined to generate tools that are precise and refined enough to support the exploration of historical texts. In addition to plain text, all kinds of corpus annotation could be exploited, such as metadata on date of publication (or more precisely: conception), author, etc.

Keeping track of diachronic form evolution has features in common with paraphrase or synonymy detection and can thus be seen as a variant of translation. Therefore the approach proposed here shares some features with work on crosslanguage information retrieval (CLIR), and the therefore applicability of of CLIRmethods such as decribed in [7] should be investigated.

Moreover, as dictionaries have been demonstrated to be useful resources for automatic word sense disambiguation[1], we foresee a role for parsed entries from historical dictionaries in this context as well. Note that the approach oulined here could also be seen as a contribution to the construction of a diachronic WordNet.[2]

In this paper we will present some experimental results for the dating of contemporary news articles. The purpose is to illustrate the potential role that temporal profiles can play in the automatic annotation of texts. The choice for contemporary content is simply given in by practical considerations. The development of statistical language models requires the availability of a huge digitised reference corpus, and as digitisation of historical text corpora has only just begun.[3] The application of temporal language models to historical data is future work.

3 Related research and background

To assist the disclosure of historical documents, time needs to be modeled somehow by the system. There is extensive literature on automatic classification of texts: There are links with work on automatic thesaurus discovery [4] and other approaches to derive synonyms and concept hierarchies from text [9, 12, 16, 18]). Temporal modeling of text does not play a role in these approaches.

Time stamps of documents have been used in several studies for browsing document collections [1,8]. In these studies, temporal metadata such as publication date are used to present or visualise the temporal structure of media collections, or just those parts that are relevant to a query. Temporal metadata can help users to zoom in on a certain period, for instance because they expect it to be more relevant for a certain topic or event than other periods. Swan and Jensen [17] were among the first to investigate which kind of statistical models were appropriate for modeling of the temporal dimension of term usage. They used simple contingency tables, similar to the wordtime matrix we introduce in the next section. In subsequent studies by Li and Croft [11] and Diaz and Jones [2], *statistical language models* are used for temporal models of term usage.

Statistical language models are simple models of language use, which were pioneered by researchers developing automatic speech recognition systems [15] and later taken up in the field of text retrieval [5,14]. A language model assigns a probability to a piece of text. Typically, we would expect a statistical language model for English to assign a much lower probability to the phrase 'mice zoo meat queue' than to the phrase 'nice to meet you'. On the basis of such probabilities a speech recognition system can be supported in picking the most likely combination of words. In this paper, we will build a similar language model for the dating task. The model assigns a probability to pieces of text, given a particular time frame.

For reasons of simplicity, instead of sophisticated *n*-gram models, representing the probabilities of phrases or word sequences up to length n, we will only use unigram models here, i.e., *n*-gram models with $n = 1$. This type of language models is commonly used in information retrieval settings and captures word or term probabilities instead of sequence probabilities. Given a certain document D, the probability to encounter a word w is calculated by the frequency of w occurring in

[1] Cf. First experiments were reported already in [10]

[2] The WordNet lexical database is a machinereadable thesaurus and semantic network developed and maintained by the Cognitive Science Laboratory at Princeton University. [3].

[3] Cf. for example projects at the Dutch National Royal Library(http://kranten.kb.nl) and the British Library (http://www.bl.uk/catalogues/newspapers/intro.asp).

D divided by the total number of words in *D*. Hence, an unigram language model of any document can be represented by a table of word frequencies.

There are several standard measures for comparing two language models, such as crossentropy or Kullback-Leibler divergence between the models. In this paper we use a normalised variant proposed by Kraaij [6], the socalled normalised log-likelihood ratio measure (NLLR) between a model *Q* and a model *D*, usually representing the query and a document to which the query is compared. *C* is a background model that is estimated on the entire collection.

$$NLLR(Q|D) = \sum_{w \in Q} P(w|Q) * \log \left(\frac{P(w|D)}{P(w|C)} \right)$$

It is easy to see that this measure is not useful if one of the terms from model *Q* is assigned zero probability by model *D*, i.e., if *P (w D) = o*, the logarithm is undefined.[4] This may occur in the case of 'unseen events', such as the absence of a word like *pilgrim* in a certain time span, *e.g.*, 1920–1930. This would normally lead to probability *P*(pilgrim *D*=1920-1930) = 0. To avoid this effect, a socalled smoothing method can be applied in estimating the probabilities of a model. In this example, smoothing would assign a very small (nonzero) probability to *pilgrim* in the time span 1920-1930.

Compared to typical document language models, temporal language models are rather large. Therefore, the issue of smoothing will probably be less important. It might even flatten out important characteristics of a specific time span. In this study we experimented with two smoothing approaches: socalled linear interpolation smoothing used by Kraaij [6] and Dirichlet smoothing [19]. In contrast to linear interpolation smoothing, the effect of Dirichletsmoothing depends directly on the size of the smoothed model. Thus, for temporal language models built from large fractions of the reference corpus the smoothing is negligible, while it is effective for models built on a small fraction of the corpus.

4 Dating of text

An example application of temporal language models is the dating of texts. In this paper we would like to show how statistical language models can be used for this task. A more precise definition of the dating task is given below:

Task definition given a datetagged reference corpus, consisting of documents from a certain time span, and a document X with unknown date within the same time span, the system should classify X according to time partitions of predefined granularity.

4.1 Reference corpus

The task definition mentions a reference corpus. Such a collection of documents with known publication date is necessary as a base for comparison. The temporal language models derived from the corpus are supposed to capture the characteristics of the vocabulary used within a certain period. The reference corpus must meet several requirements. It needs to

– be sufficiently large,
– have a balanced distribution over the represented time span,
– cover the same domain as the documents to be dated,
– and cover at least the period from which the undated documents originate.

The first requirement is formulated vaguely, because the required corpus size depends on several other parameters, such as the level of granularity imposed by the actual dating task. The main concern, however, is that a sparse data set may cause the language models to be determined by specific document characteristics rather than by temporal patterns. For similar reasons, a balanced distribution of the corpus documents over the complete time span is needed. If one temporal language model is aggregated from a few documents, while another one is based on half of the corpus, the former may suffer from data sparsity, whereas the latter may be overtrained, resulting in a model hardly distinguishable from the background collection characteristics. But even these two requirements do not help if the corpus domain differs from the topic domain of the document to be dated. Obviously, we cannot date a sports article from a newspaper with a model trained on a corpus with personal correspondences from writers. Reliable dating also requires that the publication date of target articles is covered by the reference corpus.

4 *P (w|D)* is the probability that w shows up given D

4.2 Time partitioning

In the Task Definition above there is mentioning of the granularity of the date classification. In fact, a document is not dated precisely, but our method just outputs the time span from which the document most probably originates. The reference corpus is therefore partitioned into smaller sets of documents, corresponding to time spans of the desired granularity, and the document is compared with all these time partitions. Only in case the granularity selected corresponds to temporal units of one day – in most cases not a reasonable choice – the classification will involve exact dating.

Formally the partitioning can be described as follows: if we divide n time marks t_i in ascending order over the full time span of the corpus, such that t_o marks the start of the corpus period, then each pair of adjacent time marks ($t_i < t_{i+1}$) defines a time partition C_i of corpus documents. D_j denotes any document from the corpus and $\tau(D_j)$ it's date.

$$C_i = \{D_j | t_i \leq \tau(D_j) < t_{i+1}\}.$$

We also need to distinguish between output granularity and model granularity. Whereas the first will be determined by external factors – need for publication year, week, day, etc.,– a finer model granularity can be chosen as well. Suppose we like to classify newspaper articles for the year of publication. If we build temporal language models for news on a yearly scale, we won't get very characteristic time patterns. Specific topics usually are discussed during shorter time spans and within a year almost any topic can be mentioned. An alternative would be to build models on a smaller scale, *e.g.*, weekly, while still letting the system produce labels for the year of publication. In general, as model granularity any nonoverlapping partitioning can be chosen that is finer than or equal to the one wanted for the final output.

Within this section, time span partitioning is defined as a division in nonoverlapping sections. However, in the newspaper case, the period in which a certain topic occurs in the news will not always coincide with a certain time partitioning. Normally the partition time marks t_i are equally divided over the corpus time span, resulting in an arbitrary crossing of such topic boundaries. A simple way to avoid this is to abandon the principle of overlapfree partitioning and to generate overlapping partitions with a time window moving in small steps (window size $w \leq s$ step length) over the corpus time span. For reasons of simplicity, though, we will stick here to the overlap-free partitioning.

4.3 Classification approaches for the dating task

The basic idea is to compute a language model from the undated document and compare it to the language models built from the reference corpus. Comparison of language models is an often used technique, not only for the 'classic' information retrieval task to find documents similar to a query model, but also for classification of documents, for instance in case of topic detection. In the following, we describe two such approaches for date classification. They differ in the way the language models for the reference corpus are built.

Method A: Comparison on document level The first approach is based on the idea of Diaz and Jones for temporal query profiles [2]. In a first step, all corpus documents are ranked according to their language model divergence to the undated document X, i.e., by the normalized log-likelihood ratio $NLLR(X|Dj), Dj \in C$. In a second step, a temporal profile of X is built from the set S of the top-n ranked documents by aggregating the sum of scores belonging to each time partition:

$$val(C_i) = \sum_{D_j} NLLR(X|D_j),\ D_j \in S \cap C_i.$$

The interpretation of the computed values is obvious: the higher the score of a time partition, the higher the probability that the document originates from its time span. Thus, the time partition with the highest value $val(C_i)$ is the best candidate to determine $\tau(X)$. Or in other words, the most likely publication period.

Method B: Comparison on partition level An alternative approach is to perform the aggregation beforehand by building temporal language models for each time partition, which requires to sum up the word frequencies from all documents belonging to a time partition:

$$|w \in C_i| = \sum_{D_j \in C_i} |w \in D_j|.$$

The next section discusses the building of temporal language models in more detail. Having language models for all time partitions, we are able to compare them directly with the language model for undated

documents. The highest ranked time partition C_i then determines the systems output for $\tau(X)$. This is either the time span of C_i itself, or if model and output granularity differ from each other, the enclosing time span of the output partitioning.

4.4 Data structures for temporal language models

In the previous section it was explained how aggregated temporal language models can be used for the dating task. Here we will describe two simple data structures to maintain such language models.

Figure 1. Data Structures: Table vs. Matrix Design 5

Word	Partition	Freq.
gulden	2000	1498
gulden	2001	1615
gulden	2002	481
euro	2000	10339
euro	2001	13625
euro	2002	26905
toekomst	2000	7360
toekomst	2001	6962
toekomst	2002	5141

(a) Table

	2000	2001	2002
gulden	1498	1615	481
euro	10339	13625	26905
toekomst	7360	6962	5141

(b) Matrix

Instead of storing each temporal language model separately, which introduces organizational overhead, all frequency counts can be gathered in one data structure. We will discuss two options, which are visualized in figure 1:

– a 3-attribute database table of the form [term-id, partition-id, frequency],
– or an 2-dimensional array with terms and partitions as its dimensions (wordtime matrix).

The table design is more appropriate in case of a fine time granularity or sparse reference data, because it avoids to store zero frequencies, which occurs often in these settings. If the table is sorted on term identifiers, we can also very efficiently perform a ranking of all partitions (the procedure then equals the one of a complete collection ranking on an inverted index structure).

The matrix structure on the other hand enables fast and direct positional access to all fields, even without storing redundant term and partition identifiers.

4.5 Confidence in dating

Unless we work on a very specific but timecharacteristic document type, we cannot expect that dating based on pure statistical comparison of word frequencies delivers excellent results. It is easy to imagine that an historical source and a later secondary reference share a lot of common vocabulary, although they originate from differentperiods. For the dating approach described here this is reflected by the scores for the top ranked partitions. They tend to very close, even if the corresponding time spans are scattered over the entire corpus range. This observation suggests that the dating system might be able to decide itself how confident it is about the suggested classification. This is interesting because in general reliable confidence measures for statistical tasks highly increase the usability of systems applying them. A simple confidence measure for dating could be the relative distance between the score of the topranked time partition to the scores of the following ones. A more sophisticated measure could also take into account the level of timely scattering in the topranked partitions.

5 Empirical tests on newspaper data

In order to illustrate the role of temporal language models and to assess their usefulness for dating techniques, we carried out some preliminary experiments. This section will describe the test corpus, the experiments, and the results.

5.1 The reference corpus

The reference corpus for our tests consisted of articles from two wellknown Dutch newspapers, *De Volkskrant* and *Algemeen Dagblad*, from the time span ranging from January 1999 till February 2005, in total almost 2 GB of text material.[5] Newspaper articles represent a specific document genre, but such a corpus is hetero-

5 The data stem from the socalled TwenteCorpus [13].

geneous in terms of topicality.

Indexing the corpus showed that our term vocabulary had a size of approximately 1.3 million differentword forms, including unfortunately a large fraction of spelling mistakes. To reduce the vocabulary size, we simply neglected all words occurring less than 10 times in the whole corpus, a suitable threshold to cut out a large number of spelling mistakes. Furthermore we applied a Dutch stemming algorithm, a rulebased system to reduce words to their stems. Both techniques together reduced the total vocabulary size to 170.000 words.

5.2 Dating experiments

As a set of test documents to be dated, we chose other Dutch newspapers, *Trouw, Het Parool* and *NRC Handelsblad*, originating from the same time span as the reference corpus, and let the system pick a random – thereafter fixed – sample of in total 500 articles.

Dating method A which compares models on document level, was tested on this sample, varying in the number (n) of topranked documents used to aggregate the temporal profile. Because pretests indicated that small values of n are in general beneficial, we only tested cases with $n = \{1, 10\}$. In fact, choosing $n = 1$ is equivalent to codating a text with the most similar corpus document.

To test dating method B, which compares on the model level, we built four word-time matrices differing in time granularity, ranging from a rough partitioning in quarteryears down to a granularity of two days. Only in the latter case, we tried the idea of overlapping time partitions by moving a 4-day window in 2-day steps over the corpus time span. In order to keep the results comparable, the output granularity was kept throughout all experiments to quarteryears.

The experimental results presented in the next section will show the dating performance of method B (comparison at model level) for two differentsmoothing techniques: linear interpolation smoothing and Dirichlet-smoothing. In both cases the smoothing parameters (λ, μ) are set in such a way that smoothing effects remain minimal. For the document-based dating techniques we used only linear interpolation smoothing, but here with a higher value for λ. Though in general the documents are short, their length is stable compared to the size of temporal language models, so Dirichlet-smoothing was irrelevant here.

Finally, we also tested the expressiveness of the suggested dating confidence measure. The computed value,

$$conf(\tau(X)) = -\left(\log \frac{score(C_j)}{score(C_i)}\right)$$

with C_i , C_j being the first and second ranked temporal language model, just reflects the idea to use the distance of the best scored time partition compared to the following as a confidence measure for the dating task. Within the tests we wanted to see whether the dating performance improves when a certain confidence threshold is required.

5.3 Results

Figure 2(a) shows the results of the dating experiments depending on model granularity. It displays the percentage of documents in the test set which were dated correctly. The last two bars represent the results using the direct document comparison of method A.

As said earlier, the proposed dating methods are based on word frequencies only and come with a certain error rate. However, we also see that temporal language models are far from being meaningless. Notice, that with 25 output partitions a random algorithm would only date 4% correctly.

Furthermore Figure 2(a) makes clear that method A outperforms method B. An explanation for this might be that especially in the news domain articles reporting about the same event are likely to occur, and often in highly similar wordings. Therefore the similarity between an undated document and a reference document model tends to be much higher than between the undated document and a topically unbound temporal language model.

A second observation concerns the model granularity. Apparently it pays to work with the smallest possible time spans. This allows to interpret the superiority of method A in another way: the direct document comparison can also be regarded as working with the smallest possible time unit.

For the two smoothing techniques compared, only marginal differences were found. The result comes less as a surprise, since our newspaper corpus has a relatively balanced temporal distribution, so there is less room for positive effects from Dirichletsmoothing.

Figure 2(b) shows that a higher confidence value correlates indeed with a better dating performance. Hence it can be concluded that *conf* is a useful additional metric. Still, even with a high confidence threshold, the dating reliability stays below 75%. Another im-

portant problem, not visible in the figures, is that the fraction of dating experiments gaining a confidence measure above the required threshold is decreasing rapidly. The correctly dated documents (almost 70%) at the left, are taken from the small fraction of dating experiments (approx. 10%) with a high confidence level. Here we also find an explanation for the observation that with courser model granularity more can be gained from using confidence measures. In fact, the filter effect just turned out to be higher in this case. So it is questionable whether the results on different-granularities should be compared at all.

6 Conclusion and future work

We have shown that simple statistical methods can be used to model time in a large newspaper corpus. On the basis of the system for dating texts one of the next steps will be to develop techniques that can help to link modern Dutch queries to their historical equivalents and thereby support a wide group of users with an interest in historical textual collections and the objects linked to them.

There are several issue that could be explored further in future research. Time-word matrices such as Fig. 1(b) have a property that deserves some attention. Whereas the longitudinal sections are exactly the described temporal language models for time partitions, the cross sections can be interpreted as temporal word profiles, showing the usage pattern of a certain word over the corpus time span. Although we won't use temporal *word profiles* for the dating task, they can provide interesting information about language change. Computing the time frequency correlation on all word profiles, for instance, allows to search for the most trendy words. We could also extract highly timespecific words, by searching for word profiles with just one outstanding frequency peak. Finally, exploiting the full richness of n-gram models for more refined temporal models is part of our research agenda.

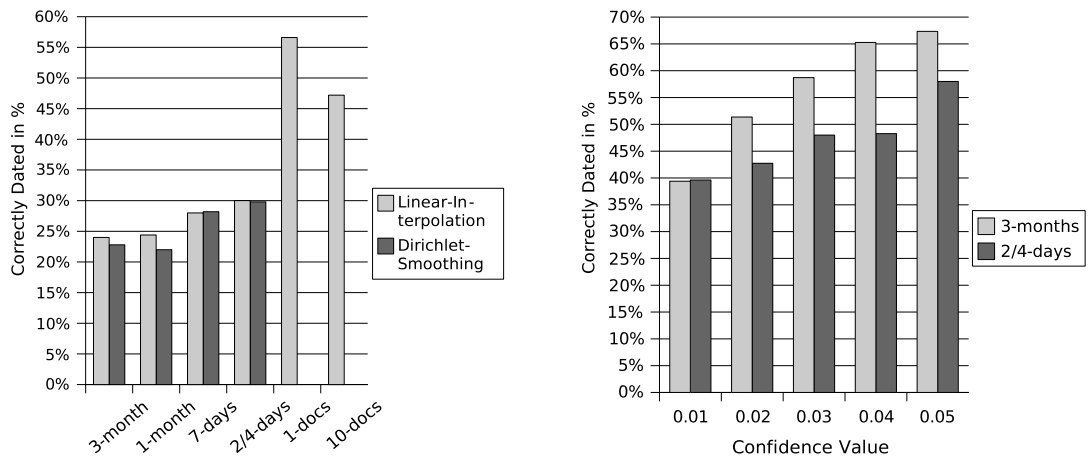
Acknowledgement

Part of the work on which this paper is based was funded by the BSIK programme MultimediaN (http://www.multimedian.nl/)'

References

[1] C. Ahlber, B. Shneiderman.'Visual information seeking: tight coupling of dynamic query filters with starfield displays'. *Proceedings of the ACM Conference on Human factors computing systems (SIGCHI)*, pages 313 – 317. 1994.

[2] F. Diaz, R. Jones. 'Using Temporal Profiles of Queries for Precision Prediction'. M. Sanderson, et al. (eds.), *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development Information Retrieval*, pages 18–24. ACM,



(a) Performance of model granularity

(b) Performance at confidence level

Figure 2. Overview dating results

Sheffield, UK, 2004.
[3] C. Fellbaum, *WordNet: an electronic lexical database*. Speech, and Communication Series. MIT Press, 1998.
[4] G. Grefenstette. *Explorations Automatic Thesaurus Discovery*. Kluwer Academic Publishers, 1994.
[5] D. Hiemstra. 'A Linguistically Motivated Probabilistic Model of Information Retrieval'. *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, pages 569–584. 1998.
[6] W. Kraaij. *Variations on language modeling for information retrieval*. Ph.D. thesis, University of Twente, Netherlands, 2004.
[7] W. Kraaij, F. de Jong. 'Transitive probabilistic CLIR models'. *Proceedings of RIAO 2004: Recherche d'Informations Assistee par Ordinateur*. 2004.
[8] V. Kumar, R. Furuta, R. Allen. 'Metadata Visualization for Digital Libraries: Interactive Timeline Editing and Review'. *Proceedings of the 3rd ACM conference on Digital libraries*, pages 126–133. 1998.
[9] D. Lawrie, W. Croft. 'Discovering and comparing topic hierarchies'. *Proceedings of RIAO*. 2000.
[10] Lesk M. 'Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a Pine Cone from an Ice Cream Cone'. *Proceedings of ACM SIGDOC*, pages. 2426, 1986.
[11] X. Li, W. Croft.'Time-Based Language Models'. *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, pages 469–475. 2003.
[12] D. Lin. 'Automatic retrieval and clustering of similar words'. *Proceedings of COLING/ACL*. 1998.
[13] R. Ordelman. *Dutch Speech Recognition Multimedia Information Retrieval*. PhD Thesis, University of Twente, Enschede, 2003.
[14] J. Ponte, W. Croft. 'A Language Modeling Approach to Information Retrieval'. *Proceedings of the 21st ACM Conference on Research and Development Information Retrieval* (SIGIR'98), pages 275–281. 1998.
[15] L. Rabiner. 'A tutorial on hidden Markov models and selected applications speech recognition'. A. Waibel, K. Lee (eds.), *Readings speech recognition*, pages 267–296. Morgan Kaufmann, 1990.
[16] M. Sanderson, W. Croft.'Deriving concept hierarchies from text'. *Proceedings of ACM SIGIR Conference on Research and Advancements Information Retrieval*. 1999.
[17] R. Swan, D. Jensen. 'Constructing Topic-Specific Timelines with Statistical Models of Word Usage'. P*roceedings of the 6th ACM Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 73–80. 2000.
[18] P. Turney. 'Mining the Web for synonyms'. P*roceedings of the Twelth European Conference on Machine Learning (ECML)*, pages 491–502. 2001.
[19] C. Zhai, J. D. Lafferty. 'A Study of Smoothing Methods for Language Models Applied to Information Retrieval'. *ACM Trans. Inf. Syst*. 22(2): 179–214, 2004.

Language change and historical inquiry

*Patrick Juola**

Language, being a product of human culture, can be one of the best indicators of human culture for historical analysis. It can show not only what a person thinks, but how she thinks, and what the tools are that she uses in her thinking. Unfortunately, language produced by one person shows only what a single person thinks, making it difficult to draw broad conclusions. Information theory may provide some methods, drawn from quantitative studies of language, for investigating large-scale historical questions based on language modelling.

After some methodological preliminaries, this paper discusses the question of language change, its causes, and the relationship between technological progress and language change. Naively, one expects that rapid changes in technology will drive (or be driven by) broad social change, which will in turn be reflected in language change. This can be confirmed by a close analysis of the language used to write the *National Geographic* magazine, and of its relationship to US patent data. The data confirms this relationship and further suggests that technology drives social change instead of reverse.

This method can also be applied to the investigation of large-scale events, or of significant personal events within the writings of a single person. This type of quantitative linguistics can provide a useful adjunct to the more traditional historical method of close reading as an analysis technique.

As any scholar struggling with a centuries-old document can attest, language changes. Depending upon the circumstances, this change in language can reflect, indicate, or perhaps even cause cultural change. Language patterns can reflect the development and rising of new power groups, of shifting patterns of influence, and of new ideas suddenly becoming widespread.

What can be learned about the history of a time by studying how language changes over that time? In particular, can the dynamics of language change - whether it is fast or slow, and what particular aspects of language show the change - be tied, even in theory, to questions of interest to a historian? Can evidence from language change be used to bolster an argument, or to suggest promising lines of inquiry? Is it even possible to measure the dynamics of language change?

To examine these questions, it is necessary, first, to understand something of the nature of language change, both in the insider's view of a linguïst and from the outsider's view of an anthropölogist or historian. We draw lightly from these traditions as a justification for a more formal, measured, and mathematical view. Using information theory as a basis, language change can be accurately measured and the measurements used as the basis for historical inquiry.

At first glance, measuring language change may appear to be a fool's errand, given the wide variety of ways in which language can vary and the difficulty of combining these ways. Just as a simple example, linguists will traditionally distinguish among various levels of language, ranging from the lexical (the words used), the morphological, the syntactic, through wide-ranging semantic and pragmatic issues. Even mere orthog-

*Duquesne University, Pittsburgh, PA 15282 USA, Tel: 412-396-5685, juola@mathcs.duq.edu

raphy can illustrate a wider historical issue, such as the increasing dominance of US over the UK in twentieth-century science, as illustrated by the 1990 acceptance of the spelling *sulfur* (cf. *sulphur)* by the International Union of Pure and Applied Chemistry.(Stevenson, 1997)

For example, the *Oxford English Dictionary* is an encyclopediac text describing the development of new lexical forms and meanings. Similarly, (Biber et al., 1998) describe the change in the (syntactic) use of modal verbs over a several hundred year period. It would be a simple step to move from a verbally descriptive to a numerical model and to give the rate of change of modal verb use.

It's not necessary to invoke four hundred years of time in order to notice language change. (Johnson, 1996) presents examples of significant change in lexical use that have occurred in only sixty years. Even more interesting are her comments on some causes of these changes:

Urbanization, industrialization, and technological advances have produced changes in occu-pation and in the implements used in the workplace and the home, which have led to changes in vocabulary.... Questions about farming, in particular, more frequently elicited 'No Response' in the 1990 interviews, as the number of farms in the South declined from 2.1 million in 1950 to 722,000 in 1975. Thus, as familiarity with farming declined, the number of speakers who admitted to lexical gaps in that domain increased.[1]

The examples of such words that she cites include: calls to cows in pasture, corn cribs, and rail fences, but also window shades and attics, which are hardly exclusive to farming. In addition to the changes driven by technology, she also discusses possible lexical effects of changes in the local economy (with increased trading and decreased economic autonomy), education (increased on average by 5.1 years between 1940-1980), and the availability of information via the media.

From a historical standpoint, two questions arise. First, the difference between modal verb use and calls to cows in pasture is illustrative. It is easy, perhaps too easy, to attribute the change in use of special cow-calls to a cultural loss of familiarity with farming, but much harder to provide any simple, clear, convincing explanation of a change in modal verb use – or even a loss of vocabulary for windowshades (which are still, after all, present in large city apartments). How should language variation without an easy explanation be interpreted? The second issue is the question of how important lexical variation is as a component of overall language dynamics? Is it possible to measure overall language change without focusing on a specific level or aspect?

One promising approach to achieving this kind of global measure lies in the mathematics of information theory (Shannon, 1948; Shannon, 1951). 'Information,' in Shannon's definition, is simply the inverse of unpredictability, or is what allows people or systems to make accurate predictions about the world. A simple coin flip will let one predict anything with 50/50 accuracy, but better knowledge can improve one's chances of correct prediction. A simple children's game may help to illustrate this. 'I'm thinking of a person.' By asking a few yes/no questions, you have to determine whom. A skillfully chosen question can eliminate as many as half the possibilities – an obvious best-possible result. If the set of possible people was 32 (2^5), it would take at most five questions to halve the set of possibilities down to a single candidate. We can thus say that choosing one item out of 32 involves a maximum of five yes/no questions (or 'bits,'in technical phrasing) of information. Similarly, choosing one person from the eight or so billion people who have ever lived should not require more than thirty-three bits of information.

Here, however, psychology and skill in asking questions begins to play a role. Although one can in theory choose any person, living or dead, from any point in history, in practical terms, one will only choose people one is familiar with. Some people are thus more probable choices than others. I can name the mayor of my own city, but not the mayor of a similar city several hundred miles away. As a U.S. citizen and resident, I am likely to know more of the names more of past presidents of the United States than of the past ministers of the Netherlands. Of the eight billion people who have ever lived, I cannot name more than a million, and probably much fewer. Using these observations as guidelines may allow you to more quickly determine the my chosen persn. Assumptions like these, if made correctly, can reduce the overall unpredictabil-

1 (Johnson, 1996) [p. 86]

ity – while, if made incorrectly, can actually increase it (perhaps I'm a doctoral student in Dutch history and you didn't realize it).

In (Shannon, 1951), Shannon illustrated how these ideas can be applied to language. Consider the following sequence of characters: 'THEREISNOREVERSEONAMOTO'. Most English speakers would guess that the next letter is more likely to be an R than a Q. Although choosing from a set of twenty-six uniformly and independently chosen symbols requires nearly five bits per symbol, the phonology, syntax, semantics, and pragmatics of English – our linguistic 'information' – can be shown to reduce our average uncertainty to approximately two bits per letter (Shannon, 1951; Schneier, 1996; Brown et al., 1992). The details of mathematical development, under the term 'entropy,' can be found in (Shannon, 1948;

Khinchin, 1957; Li and Vitanyi, 1997). Informally, we can apply the statistics of one document as a guideline to predicting a second. The quality of prediction yields a mathematical 'divergence', just like a distance one would measure with a ruler – distance is never negative; the distance between any object and itself is zero, and so forth. Most importantly, two objects that are more noticeably distinct have greater distance between the.

This distance can be used to measure the quality (or lack thereof) of the model used. If one sample is a good model of a second sample, we can say those two samples are 'close' in linguistic distance. If the measurements of model quality can be made sufBciently sensitive to work with small texts, tests like this may be practical. (Juola, 2003) describes such a test [based on the methods of (Farach et al., 1995; Wyner, 1996)] and shows how it can be generate a useful 'distance' with wide application.

In order to take meaningful measurements of language change over time, it is necessary to have suitable samples of language situated in time. This task is surprisingly difficult. First, the samples need to be comparable in a meaningful sense – otherwise, systematic differences in style or authorship may dominate subtle temporal effects. Second, the samples need to be accurately dateable, and cover a sufficient amount of time to allow useful measurements. Finally, analyses such as the present study require machine-readable, or at least machine-transcribable, corpora. And for those of us on a budget, affordability is always a nice feature. Fortunately, many periodicals, among them the *National Geographic,* which have been made available as machine-readable collections on sets of CD-ROMs at quite reasonable prices (c. US$100) that fulfil the criteria.

The first hypothesis to be tested is simply the claim that language change is detectable. Underlying this hypothesis are several assumptions, among them that language itself changes and that the changes are undirected, but cumulative. If we take English as written [in *NG*] in 1950 as a baseline, language in 1955 is likely to be different, and language in 1960 is likely to be different from the 1955 sample, and even *further* removed from our baseline of 1950. As time passes, we expect two samples of language to be more different (as measured via information theory) the greater their temporal separation. This constitutes a testable (and falsifiable) prediction.

A second hypothesis follows almost immediately from the first; if a significant correlation of linguistic with temporal separation can be established, it is reasonable to start curve-fitting. Here one can apply the notion of 'rate of change'; if points at five years distance display *X* bits of linguistic separation, while points at ten years distance display *2X* bits, then (naively), the rate of change is *2X/10* bits per year. One can easily test whether the rate of language change is itself uniform (another falsifiable hypothesis) over comparable time scales.

Both of these hypotheses have been tested (Juola, 2003) (and results follow below). For every pair of distinct documents, the linguistic difference between the two was computed, as was the number of years of separation. Analysis was performed over decades[2] and a 'rate of change' calculated.

In all cases, the expected results were obtained. A sample plot is presented in figure 1, covering the 1949-1958 period. As can be seen, documents of the same year were typically closer than documents of distant years. This can be more formally shown by the fitted (dotted) regression. Table 1 shows the values of the regression slopes for all studied decades. As expected, all slopes are positive (significantly so in all cases but the 1940s) reflecting measurable change in

2 Pseudo-decades. The period herewith referred to as 'the 1940s' is actually the period 1939-1948; the period referred to as 'the 1950s' is 1949-1958, and so forth. *Caveat lector.*

Table 1. Average rates of language change for various periods

Period	Average rate
1939-1948	0.0039 bits/year (* p = 0.0741)
1949-1958	0.0178 bits/year
1959-1968	0.0167 bits/year
1969-1978	0.0111 bits/year

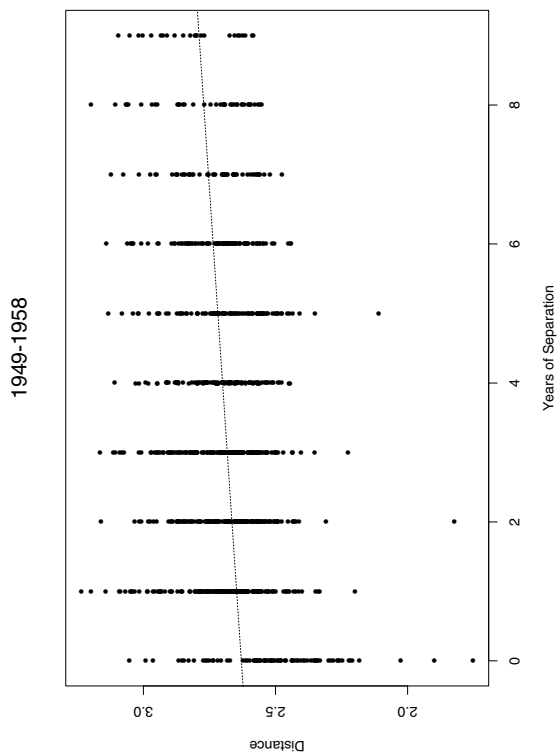


Figure 1. Temporal distance vs. measured linguistic distance, with regression line

language over a single decade. The first major result, then, is that language change does occur, is detectable by the techniques described above, and that linguistic diversity increases, rather than decreases, over time.

Over the period of 1949-1958, language changed at an average measured rate of 0.0178 bits/year. In practical terms, then, a person with perfect knowledge of English who had fallen into a coma in 1949 could have awakened in January 1958, still with a very good practical knowledge of English. However, his background would have been sufficiently out-of-date that he would have played Shannon's language game relatively badly, averaging about 0.18 questions per letter poorer than his less well-rested contemporaries, reflecting his lack of knowledge of 'current' concerns and idioms. For example, 'the Pill' (a technology under development in the 1950s and a phrase coined, according to the *OED,* in 1957), would probably be meaningless to him.

It is also observed that the measured rate of change for the various decades differs, in most cases significantly so. One can easily observe that the 1940s had less change (significantly so), than the 1970s, the 1970s had significantly less change than the 1950s and 1960s, and, of course, the 1940s were significantly smaller than the 1950s/60s. The difference between the rates of change of the 1950s and 1960s was not significant, although it may suggest directions for future and more elaborate experiments.

Observing merely that language changes at different rate may itself be of little interest unless an explanation can be found. One possible explanation often given for language change is technological progress; as new inventions become available, people have new things to talk about. Alternatively, as culture changes, people have new desires which may be reflected in the kinds of technology that is developed. If either of these hypotheses were correct, we would expect to see a correlation between the rate of language change as measured above, and the rate of technological progress. Furthermore, it may be possible to infer the causal link — does technology drive language change or does culture/language drive technological progress, by observing if rapid change in language leads or lags rapid technological change.

To perform this experiment, data from the USPTO and the Census was used to determine a per-capita rate of patents granted (in the United States) over the period 1930-1980. For each year 1939-1978, the 'instantaneous' rate of language change was obtained by averaging the measured changes over a five-year window. Although correlating these sequences as-is yielded no significant relationship, a significant correlation ($r > 0.32$, $p < 0.05$) was obtained between language change and the patent rate about *eight years earlier.* (Significance is actually achieved at any lag between 6-10 years). It is tempting to conclude that technology

can be a significant factor in language change, but that it takes nearly ten years for the effect of technological innovation to show up in the language of general readership magazines.

This experiment has been partially replicated (Juola, 2003) in studies involving different corpora. The Historic Pittsburgh project, a joint project of the University of Pittsburgh Digital Research Library and the

Historical Society of Western Pennsylvania, has been digitizing and storing materials of historical interest for Pittsburgh and western Pennsylvania. Included in their materials is the Full-Text Collection, a set of newly electronicized non-fiction and reference material published in the 19th and early 20th century covering, according to their project description, 'the growth and development of Pittsburgh and the surrounding Western Pennsylvania area from the period of exploration and settlement to the period of industrial revolution and modernization.' From this collect, 68 books of appropriate style, published between 1900 and 1939, were selected. These were broken down by decade and analyzed for the rate of change, as before.

Omitting statistical detail in these findings, similar patterns emerged; table 2 shows that (with one exception, perilously close to zero and within the bounds of noise), all changes were positive, but that the rate varied from period to period.

The Historic Pittsburgh documents are unfortunately not directly comparable to the *National Geographic* texts. Not only do they come from an altogether different period, but they lack the close editing, editorial consistency, and commonalities in staff authorship. In addition, the dating of individual documents is problematic; a book published in 1920 may well have been written in drafts since the Civil War. Despite these differences, similar patterns of change can be observed, suggesting that these patterns are *not* the result of (independent) artifacts in two entirely dissimilar sets of data.

A legitimate counter-suggestion might be that we are simply measuring lexical change and not in fact the overall change theory describes. Especially considering the proposed relationship between technological and linguistic change, might not people simply have new things – new words – to talk about? This has been tested in another series of experiments (Juola, 2002a; Juola, 2002b) comparing the degree of lexical difference, the percentage of types (or alternatively of tokens) that appear in one document but not in another. If language change is only, or even primarily, lexical change, then a similar analysis of lexical dissimilarity should reveal similar patterns.

Table 2. Average rates of language change for early 20th century

Period	Average rate
1900-1909	0.0233 bits/year
19410-1919	-0.0011 bits/year
1920-1929	0.0450 bits/year
1930-1939	0.0180 bits/year

Unfortunately, this kind of analysis requires substantially more data to get a representative sample of the lexicon, and the *NG* fragments were too short for such analysis. The novel-length samples from Historie Pittsburgh are however ideal. Using this data, we have analyzed whether or not the rate of lexical dissimilarly changes within decades, and also correlated the observed (instantaneous) overall change with the rate of lexical change.

Decade-based analysis showed no significant results, and in particular, the degree of lexical dissimilarity did not notably change as temporal difference increased. Correlation of lexical dissimilarity with the overall measured rates of changed also showed no significant results. Although it is difficult to infer anything from a lack of significant results, this strongly argues that 'mere' lexical change is not an adequate explanation of how and why language has changed over the measured periods.

We have, then, that language *does* change, and that that change can be algorithmically perceptible, even over periods as small as a decade. Even in isolation, this is a relatively important development and suggests a new addition for the toolbox of language scholars interested in quantifying language change and variation. From a sociological perspective, however, the findings that language change is not uniform over time is more immediately interesting. Given that language appeared to change relatively quickly between 1949-1968 and relatively slowly between 1939-1948, why did that occur? More broadly, what's the difference between the 1940s and the 1950s?

And, of course, a historian, faced with a question of such tremendous breadth, could only respond by writing a book (or a series). Not being a historian,

the author can offer only the most naive suggestions based on gross-scale perception of cultural pressure. For instance, the 1940s were the period of the Second World War, one of the most significant political events of the twentieth centure, especially in terms of direct, personal effects on the 'average' American. For the first time in a generation, people were dragged from home and hearth and placed in the theater of world opinion, along with several million of their countryfolk from thousands of miles away. That this event would not somehow leave its mark in the language of those millions of people is implausible.

In the course of this war, these millions of people would be exposed to new experiences, new ideas, new technologies, and sometimes an entirely new linguistic environment. Five years later, the veterans would be taking *their* new experiences home and retelling them to the people who had not gone – and the advertisers and journalists, in the new peacetime prosperity, would be writing to the veterans in a language they believed would be effective for these people, not necessarily the same language they would have used earlier. Of course, we have no reason to assume that this change would necessarily have been instantaneous, but the cumulative pressures of a billion new experiences could have caused tremendous change. If, in fact, it took more than three years after the end of the war for the 'new' language of the veterans to reach the relatively conservative *National Geographic,* one would expect tremendous change in the 1950s, relative to the 1940s. This view is supported by the numbers in the second experiment, which also show an increase in language change in the decade immediately after a major war.

From a methodological standpoint, the most important conclusion is simply: the technique described here works for measuring language change and variation. That language changes is and has been unassailable; how fast it changes has not been the subject of much agreement. This paper has demonstrated how to make a direct quantitative measurement of the amount of language difference from one document to a second, even from samples of only a few thousand characters, and yet obtain meaningful measurements.

From a linguistic, or a psycholinguistic, perspective, much additional work is necessary to explain the numeric findings. 'Yes,' one can say, 'language changed more in the 1950s than other decades.' However, what *form* did this change take? Is technologically-driven change primarily lexical, as suggested and refuted above? Is the rate of lexical innovation different from the rate of syntactic innovation? Does this represent merely a pragmatic difference in what people choose to write/talk about, or is there a fundamental difference in the representation of language going on in people's heads? In addition to requiring close analysis of the relevant documents, new techniques may need to be developed to test these conjectures. And, finally, from a historica! perspective, this may suggest a new indicator of cultural changes and perhaps a new technique to spot previously unsuspected sources for linguistic and cultural pressures. At the very least, information of this sort can be a finger pointing at new information to be read, evaluated, and explained. The current work strongly suggests that language change *is* related to technological change. However, technological change is clearly not the only factor in language change. A similar investigation could, and should, be performed on any other proposed factors that engender or hinder linguistic change. But merely by allowing language change to be accurately measured, one can use this as a tooi to unpack these components of society and examine them individually.

References

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press, Cambridge.

Brown, P. F., Della Pietra, V. J., Mercer, R. L., Della Pietra, S. A., and Lai, J. C. (1992). An estimate of an upper bound for the entropy of English. *Computational Linguistics,* 18(l):31-40.

Farach, M., Noordewier, M., Savari, S., Shepp, L., Wyner, A., and Ziv, J. (1995). On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. In *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms,* pages 48-57, San Francisco, California.

Johnson, E. (1996). *Lexical Change And Variation in the Southeastern United States 1930-1990.* University of Alabama Press, Tucaloosa, Alabama.

Juola, P. (2002a). 'lexical change as a factor in linguistic change. In *Proc. Measuring Lexical Variation and Change: A Symposium on Quantitative Sociolexicology,* Leuven, Belgium.

Juola, P. (2002b). The significance of lexical choice in language change. In *Proc. Workshop on Quantitative Investigations in Theoretical Linguistics (QITL),* Osnabruek, Germany.

Juola, P. (2003). The time course of language change. *Computers and the Humanities,* 37(l):77-96. Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory.* Dover Publications, New York.

Li, M. and Vitanyi, P. (1997). *An Introduction to Koïmogorov Complexüy and lts Applications.* Graduate Texts in Computer Science. Springer, New York, 2nd edition.

Schneier, B. (1996). *Applied Cryptography, Second Edition: Protocols, Algorithms and Source Code in C.* -John Wiley and Sons, Inc., New York.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal,* 27(4):379-423.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal,* 30(1):50-64.

Stevenson, R. (1997). Comments: An editor's lot is not always a happy one. *Chemistry International,*

Wyner, A. J. (1996). Entropy estimation and patterns. In *Proceedings of the 1996 Workshop on Information Theory.*

Not ready for the Semantic Web: A field study of subject gateways on contemporary history

Michael Kröll

1. Introduction

In the German speaking language areas, three major web-based subject gateways focusing on Contemporary History have been built up during the last ten years. These projects, essentially working in parallel, share the pretence of being a main reference of its kind. Nevertheless, they differ substantially with regard to certain project characteristics, such as their date of establishment or disposable resources.

This paper provides an overview of technical and methodological aspects experienced by undertaking a comparative analysis of
www.zeitgeschichte-online.de,
www.vl-zeitgeschichte.de and
zis.uibk.ac.at as three major examples of German-speaking web-based subject gateways on Contemporary History.

2. Basic information about the three web-based subject gateways

The 'Zeitgeschichte Information System (ZIS)', online since early 1995, is the longest-running web-based subject gateway on Contemporary History among the three projects examined. Maintained by the Institute for Contemporary History at the University of Innsbruck, its main features include an annotated link database comprising about 800 entries, primary sources of 20th century Austrian history, a documentation of the history of South Tyrol and a documentation on 'Austria & Israel since 1945'. The most recent review of ZIS is available online at (Gehringer 2003).

The 'Virtual Library Zeitgeschichte (VLZ)', part of the *W3C Virtual Library*[1], was the result of the merge of the Virtual Library sections 'Third Reich/World War II' with '20th Century' in 2003. The VLZ is managed by a team of historians on an honorary basis. Its main feature represents a link database including about 700 entries. The most recent review of VLZ can be found at (Böhler 2004).

The 'Zeitgeschichte-Online (ZOL)' project is a joint endeavor of the 'Zentrum für Zeithistorische Forschung' (ZZF), Potsdam and the 'Staatsbibliothek zu Berlin – Preußischer Kulturbesitz' (SBB), Berlin. The subject gateway went online in early 2004, and is supported in close co-operation with the two probably most important subject gateways on History in the German speaking area, 'Clio-Online'[2] and 'H-Soz-u-Kult'[3]. 'Zeitgeschichte-Online' features a database on institutions related to and persons working in the field of Contemporary History, a sub-branch of the *H-Net* list H-Soz-u-Kult called 'H-Soz-u-Kult/Zeitgeschichte', pertinent subject foci, subject related online discussion fora, and a link database including about 2,100 entries. The most recent review of ZOL is available at (Van Laak 2004). Judging from the infrastructural background of the co-operation partners, the 'Zeitgeschichte-Online' project's subject gateway should by far show the highest grade of professionalism of the three subject gateways at issue.

3. Generic technical web-page evaluation methods

The Internet could not exist without technical standards. However, in the light of the majority of web-pages currently available, one would be inclined to think that quite the opposite is true. Given the lack of conformity with regard to technical standards, a considerable lack of interoperability, accessibility, and usability can be discerned.

1 http://vlib.org/
2 http://www.clio-online.de/ One of CLIO-Online's sponsors is the potent *Deutsche Forschungsgemeinschaft* (DFG).
3 http://hsozkult.geschichte.hu-berlin.de/

3.1 Syntactic standards

The *W3-Consortium*[4], mainly responsible for creating web-standards, provides validation services for syntactic web-page standards. Using these validators[5] for the start-pages of the three subject gateways to test *HTML* and *CSS* validity has shown that all pages are invalid with error counts ranging from 6 to 410. Despite being syntactically invalid, the document will still be accessible using most browsers. It has to be concluded that the creators of the *HTML*- and *CSS*-pages simply are not aware or are not concerned about standard conformance.[6]

3.2 Metadata Standards

Only if a web-page is syntactically formalized, i.e. by being marked-up in valid *(X)HTML* can value-added processing by software tools be undertaken. Adding a formal and explicit meaning to content by using metadata is one of the cornerstones of a future Semantic Web. Implementing *Dublin Core*[7] as the de-facto metadata standard for one's web pages would be a first step towards that goal. Only one of the three subject gateways at issue, the 'Zeitgeschichte Informations System', partly[8] uses *Dublin Core* in its *HTML* pages.

3.3 Accessibility Guidelines

Another generic factor for the quality of web-pages is their conformance to accessibility guidelines like the W3C's *WAI*[9] or the U.S. Government's *Section 508*[10]. Again, the use of validators[11] to check conformance shows that none of the three subject-gateways passes the tests. In contrast to the *HTML*- and *CSS*-validation tests, the effects of non-accessible pages are far more severe for people with disabilities and therefore a strong call for action to make web-pages accessible has to be stated.

4 Setting up a framework for specific analyses

The discussed generic technical web-page evaluation methods can only provide rather generic answers. For more specific questions, *e.g.* in quantitative analysis, more specifically tailored software is needed. In the course of the comparative analysis of the three aforementioned subject gateways, a crawler program has been developed for harvesting the content of each subject gateway's link database. The crawler has to use heuristics to map the crawled data into a common database. There are two reasons for this: Firstly, none of the subject gateways offers a formalized public interface to access its databases, like for example providing a custom *Web Service*. Therefore, the link databases have to be harvested by parsing their *HTML* output. Secondly, it is necessary to map the harvested link metadata to a common scheme. None of the three subject gateways declares to use a common metadata scheme, thus a specific conceptual mapping to the *Dublin Core*-using aggregate database had to be set up. Analyzing the data to be harvested showed that these metadata mappings could not be static. In case of 'Zeitgeschichte-Online', for example, the fields 'Autor' ('author'), 'Herausgeber' ('editor'), and 'Veröffentlicht durch' ('published by') could not be mapped 1:1 to DC-Creator and/or DC-Publisher, the only two DC-fields available for matching in that case. Depending on the presence of data in one of the three fields, a different semantic meaning had to be applied. Using standardized means for providing access to one's metadata or archival information, *e.g.* by implementing an interoperable *OAI-PMH*[12] data provider interface, could avoid potential errors due to such ambiguities.

The common database storing the harvester results has been implemented using the *PostgreSQL*[13] RDBMS. The crawlers have been implemented using

4 http://www.w3.org/

5 The validator used for HTML http://validator.w3.org/, for CSS http://jigsaw.w3.org/css-validator/.

6 A comprehensive discussion of the implications thereby created, exceed the scope of this paper.

7 http://www.dublincore.org/ – A brief overview and introduction of its usage is provided by Hillman 2003.

8 Dublin Core is used on the entry pages only.

9 http://www.w3.org/WAI/

10 Implementation of Section 508 is legally binding for U.S. federal agencies. More information can be found at http://www.section508.gov/

11 Validator used for WAI/WCAG and Section 508: http://www.contentquality.com/

12 OAI PMH stands for the Open Archive Initiative Protocol for Metadata Harvesting. See Caplan 2004 for information about the protocol and Kelly 2004 for general consideration on Digital Library interoperability.

13 http://www.postgresql.org/

Perl[14] and *Perl CPAN*[15] modules. They retrieved the metadata from the contents of the link databases assignable to the *Dublin Core Metadata Element Set*, the *HTTP* status code and *MD5* checksum of the database item's content. In addition, each link database item's homepage was crawled recursively to three levels of depth, to store the out-links to other link database items.

5 Using the custom analysis framework

Having crawled 3,646 interlinked items holding a number of attributes provides copious space for analysis. In the following, a selection of options for analysis will be discussed.

5.1 Resource Identifier Validation

HTTP status codes tell us about the availability of a resource. Status codes greater than 400 denote an invalid resource, which could *inter alia* be the result of either '404 Not Found' or '500 Server Error'. The following table provides an overview of invalid items in the aggregated link database grouped by subject gateway as of April, 14th 2005.

Subject Gateway	Total Items	Invalid Items	Invalid Items %
Zeitgeschichte Informationssystem	822	178	21 %
Virtual Library Zeitgeschichte	693	66	9 %
Zeitgeschichte-Online	2,131	82	3 %

The disparate percentage of invalid items from the ZIS database could be explained by the fact that the last update of that database has been performed at August, 13th 2003. Unfortunately, the last update timestamps of the other databases could not be ascertained from their homepages.

5.2 Subject Classification Analysis

Subject categorization is one of the most challenging and time consuming tasks of metadata classification. It is also one of the tasks still most recalcitrant to automation, as a recent report on *Automated Metadata Classification* (Greenberg 2005) has shown.

Full blown taxonomies, thesauri or ontologies may be too costly for metadata-input. Still, applications should at least be designed to avoid free-form data entry as it seems to be the case with 'Zeitgeschichte-Online', where the concept of 'Arbeiterbewegung' (labor movement) can be found in the four different keywords 'Arbeiterbewegeung', 'Arbeiterbewegung', 'Arbeiterbewegungen', 'Arbeiterbewgung'. Similar spelling errors or redundant classifications can also be found in the other two databases.

The overall distribution of keyword usage in the aggregated database can be interpreted as an indication of the most popular research topics in web-present Contemporary History in the German language area:

Top 10 Keywords (Number of Occurrences)	
425	Nationalsozialismus
288	Holocaust
243	Sozialgeschichte
203	Widerstand

Looking at the keyword distribution broken down by subject gateway, a maverick can be noticed:

Subject Gateway	Total items	Distinct keywords	KWs with only one occurence /per distinct keywords	%	Most used keyword
Zeitgeschichte Infomations-system	822	166	26	15.66%	Holocaust
Virtual Library Zeit-geschichte	693	126	19	15.08%	'Drittes Reich'
Zeitgeschichte-Online	2,131	1,146	502	**43.80%**	National-sozialismus

Approximately 44% of the keywords used by 'Zeitgeschichte-Online' are used only once. This can be interpreted either as an indication for a thematically wide-

14 http://www.perl.org/
15 http://cpan.perl.org/

spread content of the link-database, or – for the worse – for a lack of stringent rules for subject classification. A third possible interpretation for that outlier is a rather practical one: 'Zeitgeschichte-Online' is aggregating link database entries from partner institutions, which represents a factor potentially increasing the content and classification diversity.

5.3 Common Resource Identifiers Analysis

When comparing three link databases with a common content focus, the question about commonly shared *URL*s seems rather obvious. In the case of the present study, the answers to that question turned out to be a surprise: From the total of 3,370 distinct normalized *URL*s, only 195 (5.79 %) are common to at least two subject gateways and only 24 *(sic!)* (0.71 %) are common to all three. Especially the last number makes a statistician doubt his methods and findings. However, the result stayed the same after double-checking.

There are some possible interpretations for that very low number of shared *URL*s. However, each interpretation only provides a partial explanation. The fact that no entry of the ZIS database is newer than August 2003 could be one factor, another that the 'Virtual Library Zeitgeschichte' has a relatively specific subject focus. The disproportionately small number of shared *URL*s questions the comparability and the authority of the three subject gateways, irrespective of the reasons for it and irrespective of the gateways' – at least original – authoritative pretence.

Shared URLs by subject gateway	Virtual Library Zeitgeschichte	Zeit-geschichte-Online
Zeitgeschichte Informationssystem	50	77
Virtual Library Zeitgeschichte	n/a	116

Having only 24 shared link database entries facilitates a classification analysis. Subsequent to the general classification analysis undertaken above, the results are not unexpected: Not a single keyword is used by all three subject gateways for one of the 24 items. For 16 of the 24 items, at least one identical keyword has been used by two of the three subject gateways. In the remaining 8 cases, all the keywords used per database item were different. As an example, the classification of the resource http://chronik-der-mauer.de/ is shown in the following table:

Zeitgeschichte Informationssystem	Berliner Mauer, Bibliographie, DDR, Kalter Krieg
Virtual Library Zeitgeschichte	Mauerbau und Grenzbefestigung, Nachkriegszeit
Zeitgeschichte-Online	Berlin, Berlinpolitik, Grenzen, Mauerbau, Zeitzeuge

It can be noticed that some of the used keywords, *e.g.* 'Berlin' and 'Grenzen' (borders) are formulated very broad and therefore very unspecific.

5.4 Duplicate Pointers Analysis

The web offers several possibilities to address the same resources under different *URL*s. For example, host aliases or directory index files allow
http://www.h-net.org/~german/ and
http://www2.h-net.msu.edu/~german/ or
http://www.icbh.ac.uk/icbh/ and
http://www.ihrinfo.ac.uk/icbh/welcome.html to point to the same documents. Identical content can be identified by using *message digest algorithms*. After harvesting the link databases of the three subject gateways at issue, the crawler stored *MD5* checksums for each document. Using that mean, it was possible to identify several documents stored under different *URL*s. The ZIS database stores 9 distinct documents under 19 different *URL*s, the ZOL database 11 under 26, and the VLZ database 4 distinct documents under 8 *URL*s. In other words, the same document had been entered under two different *URL*s into the VLZ database in four cases.

5.5 Information Network Analysis

As mentioned earlier, the crawler programs stored the out-links of the database items' web-pages to the other items' web-pages over a depth of three levels (clicks). The resulting graph can be analyzed by using a variety of methods related to the field of network analysis.

The *PageRank* algorithm (Page 1998), popularized since its use by the Google search engine, can be used to help determine a page's relevance. For our ZIS/ZOL/VLZ network, the top-five ranked *URL*s are shown in the following table:

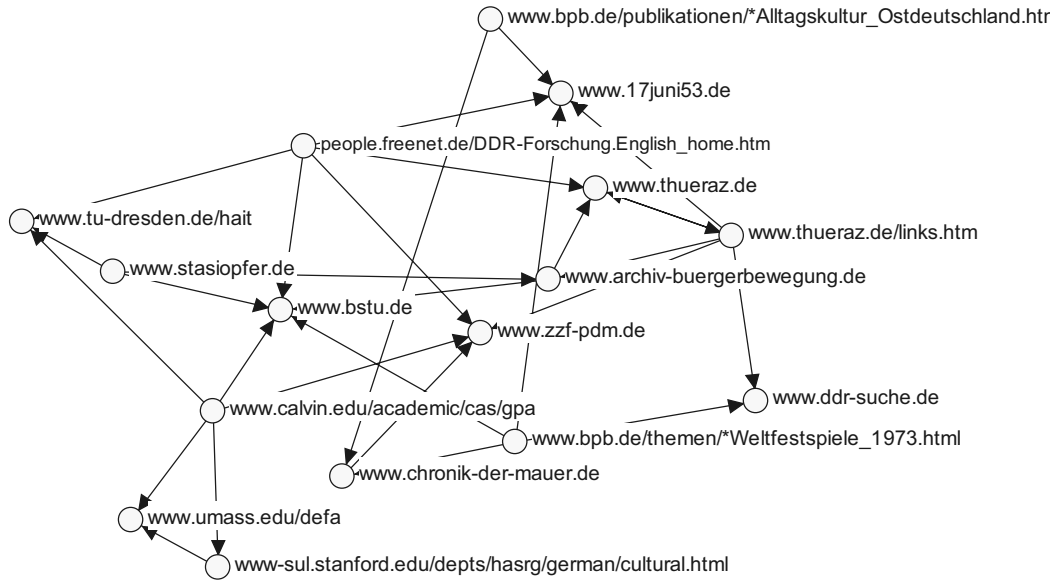


Diagram 1: Network based on keywords 'DDR' and 'Deutsche Demokratische Republik 1949-1990'. Nodes with a degree of one have been removed for better visibility.

URL http://	**In-Degree**	**PageRank** (ZIS/ZOL/ VLZ network, n=2278)
www.dhm.de/	123	8.87
www.wiesenthal.com/	109	7.94
www.ubka.uni-karlsruhe.de/kvk.html	119	7.76
www.iwm.org.uk/	24	7.60
www.iwmcollections.org.uk/	19	7.41

After converting the directed link graph to a *binary asymmetric adjacency matrix*, the wide-ranged power of network analysis software tools like *UCINET* (Borgatti 1999) or *Pajek* (Batagelj 2005) can be put to use.

Because the density of ties in the whole ZIS/ZOL/ VLZ network is very low (0.3%), we will partition the network matrix along keyword-based parameters to be able to tell about *degree*-, *betweenness*-, and *closeness centrality*, as well as other network analysis concepts like *Bonacich Power* Indices or *cliques*[16] for the subject-related sub-networks.

Using network visualization software like *NetDraw* (Borgatti 2002) allows for a quick and comprehensive overview of such sub-networks as shown in the exemplary diagram 1

At 4.73%, the density of this 'GDR-Network' is much higher compared to the overall ZIS/ZOL/VLZ network. www.bstu.de, www.17juni53.de, and www.zzfg-pdm.de can be identified as the three central web-pages by catching a glimpse of the diagram. The 'GDR-Network' only has 31 nodes, whereas the total number of distinct items sharing one of the two keywords 'DDR' or 'Deutsche Demokratische Republik 1949-1990' in the aggregate database is 59. That means that almost half of those items are not cross-linked by the others. Because of this and the overall very low density of the network, the case for web-subject gateways filling those missing links can easily be established, assuming that it is not Google[17] that will be forestalling this function in the pre-Semantic Web[18] era.

16 An introduction to Social Network Analysis Methods provides Hanneman 2001.
17 http://www.google.com/
18 An general overview on the Semantic Web is provided by Miller 2003. In the context of the discussed field study, an overview of the Semantic Web from a Web-Mining perspective is given by Berendt 2004.

6 Conclusions

Focusing on the methodological and technical aspects of a comparative analysis of three major web-based subject gateways on Contemporary History in the German language area has shown that their individual support for standards is poor. Neither syntactical, nor metadata nor accessibility standards are applied to an adequate degree. Specifically tailored crawlers had to be developed to use heuristics to harvest the data of the link databases because no standards-based interoperability framework is used by the subject gateways. Analyzing the data in the harvested aggregate link database has demonstrated that the subject gateways' management software does not prevent to enter duplicate resources and does not take measures to avoid spelling mistakes during classification.

The three subject gateways' link databases only share less than 1% of their total sum of *URL*s, which indicates that the overall network density of the ZIS/VLZ/ZOL network is very low. Applying information network analysis methods further supports this indication.

The field study has shown that the three subject gateways are not making use of the current web-technologies' and metadata-standards' potential. Especially in the light of the development towards a Semantic Web it is to hope that the subject gateways on Contemporary History at issue will improve their standard compliance and interoperability, so that they will not remain in the domain of a comparatively meaningless web.

7 References

Batagelj, V. (2005). Pajek 1.04.

Berendt, B. H., Andreas; Mladenic, Dunja; van Someren, Maarten; Spiliopoulou, Myra; Stumme, Gerd (2004). A Roadmap for Web-Mining: From Web to Semantic Web. Web Mining: From Web to Semantic Web. First European Web Mining Forum, EWMF 2003, Cavtat-Dubrovnik, Croatia.

Böhler, I. G., Michael (2004). 'Wendungen nach innen? Selektive Blicke auf die Zeitgeschichte.' Zeithistorische Forschungen/Studies in Contemporary History 1.

Borgatti, S. P. (2002). NetDraw: Graph Visualization Software, Analytic Technologies.

Borgatti, S. P. E., M.G.; Freeman, L.C. (1999). UCINET 6.0 Version 1.00, Analytic Technologies.

Caplan, P. (2004). 'OAI-PMH.' Computers in Libraries 24(2): 24.

Gehringer, H. (2003). 'Rez. WWW: Zeitgeschichte Informations System (ZIS).' Retrieved 2005-01-18, from http://hsozkult.geschichte.hu-berlin.de/rezensionen/id=18&type=rezwww.

Greenberg, J. S., Kristina; Crystal, Abe (2005). Final Report for the AMeGA (Automatic Metadata Generation Applications) Project, UNC School of Information and Library Science.

Hanneman, R. A. (2001). 'Introduction to Social Network Methods.' Retrieved 2005-04-10, from http://faculty.ucr.edu/~hanneman/SOC157/NETTEXT.PDF.

Hillman, D. (2003). 'Using Dublin Core.' Retrieved 2005-03-03, from http://www.dublincore.org/documents/usageguide/.

Kelly, B. (2004). Interoperable Digital Library Programmes? We Must Have QA! Research and Advanced Technology for Digital Libraries, 8th European Conference, ECDL 2004, Bath, UK.

Miller, E. S., Ralph (2003). 'An Overview of W3C Semantic Web Activity.' Bulletin of the American Society for Information Science and Technology 29(4): 8-11.

Page, L. B., Sergey; Motwani, Rajeev; Winograd, Terry (1998). The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project.

Van Laak, D. (2004). 'Rez. WWW: Zeitgeschichte-online.' Retrieved 2005-01-18, from http://hsozkult.geschichte.hu-berlin.de/rezensionen/id=48&type=rezwww.

Moving through the city: residential mobility and social segregation in Amsterdam 1890-1940

Henk Laloli

Theories of urban change in cities (the Chicago school, Hoyt, social area analysis) continue to give inspiration, but when specific cases are investigated they show many divergences from supposed patterns. Does Amsterdam's development conform to that of American cities or even other European cities? One idea that would not stand up to scrutiny is the radical change of the old city. Before the industrialisation Amsterdam was already divided in areas with dominant social groups. The inner city contained a mixture of people and activities. Around it a canal zone (Grachtengordel) was laid out for the rich and round these canals the working-class districts could be found. This pattern remained intact well into the 20th century. The reasons for this were several. The old city was not radically altered in terms of housing: some bad housing was tore down, but the replacements were filled with only a slightly higher income group. The elite, once living in the city centre, moved out, but not completely, and established itself in the new districts in the south or in the regional suburbs. The old city centre remained a mixed area: more homogenous in terms of population, but mixed in terms of functions used for housing, offices and shopping. The harbour and new large industries were actually quite near the old city, because they were built along the IJ the workers could stay in the same place despite job changes.

The mentioned theories of the modern city see its residential and economic functions diverge into separate zones or sectors that lead to the rise and deepening of social segregation in the city.[1] They suppose that economic and occupational change led to increasing social differentiation, individualisation and social mobility. One outcome was different residential choices of different social classes. Another would be constant filtering down of housing: the lowest income groups establishing themselves into the houses left vacant by better off groups escaping to an outer zone of new housing. These theories presuppose the unfettered reign of free markets and individual choices. Certainly, Amsterdam showed in the pre-1900 expansion all these signs of unchecked commercial housing development, but after the establishment of a Housing Act in 1901 the city and the national government had powers to regulate the free market and plan urban development. Still, it remains to be seen if the forces of economic change and social mobility were checked.

We can define social segregation as the spatial clustering of people in terms of several aspects: income, ethnicity, family size, health etc. Research into these matters has proposed that segregation takes different spatial characteristics according to different aspects: the economic differences fall out into sectors, the family differences into zones, and the ethnic into clusters. Segregation according to these characteristics can be studied by using data on income, housing, rents and health at the district level. I will also use data from individual life courses that show residential change and relate these to broader patterns. The above mentioned theories lead us to suppose that Amsterdam's residential segregation by income increased, that the poor were either chased out of the inner city or remained

1 For a critical overview of these theories see: M. Cadwallader. *Urban geography: an analytical approach.* Upper Saddle River: Prentice Hall, 1996. Chapter 6, 'Urban social areas'.

locked into a downgraded old area, that the better off workers and middle classes preferred the newly built zones to the old city. In short, that social and residential mobility led to increasing social segregation in a pattern of constant resettlement of different groups outwards.

Approaches

In geography there exists a huge volume of research on residential mobility and also on social segregation. In history we see less of this. In the study of Dutch cities some researchers have paid special attention to social segregation. Michiel Wagenaar did this for Amsterdam for the period 1878-1915 and he concluded that between these dates there was decreasing heterogeneity and increasing segregation. He compared the quarters of the city on income, industry, schooling and population density.[2] The new districts became either lower-middle class areas or elite and the poorer districts lost their elite and middle groups. The inner city lost some its elite and paupers. The canal zone remained one of the richest areas in the whole city and the remarkable thing is that many rich people did not move. The poor could not move outside the working-class districts in 1915 because rents were too high elsewhere. Wagenaar explains the increasing social segregation in the new areas by the housing rent, which was too high for most of the poorer people. He thinks that municipality was unable to change the structure of segregated development. The poor remained tied to the structures support and credit that were part of their survival in the old districts.

In a recent article Jan Kok and others investigated if the number of residential moves in Amsterdam can be interpreted as part of workers living strategies.[3] They show that the number of movements can be explained in general by the housing market. When a large number of houses was available until approximately 1910 the number of movements was the highest. After that period housing shortage grew till into the nineteen twenties and the number of movements dropped. There are clear social differences in the number of movements: casual workers are the ones that move most before 1910. The authors relate this to their living strategies. The poor used these changes as part of making ends meet means. A house with a somewhat lower rent would be taken up in no time. The poor also moved inside a small area: most movements took place inside the street and they hardly left the old city centre. This again can be explained in financial terms: housing in the newly built areas was more expensive and few unskilled workers could afford it. The support network of family and neighbours (if it existed) was nearby. The findings agree with other research for later periods: people change addresses frequently when they have a young family, there is less change in the later life course, most movements occur over a short distance, in a small sector of the city.[4]

These authors did not analyze the movements as such, the geographic beginnings and destinations and the clustering of movements in certain areas. I propose to do that. They limited themselves to the frequency of movements and they only categorized the city into five areas, whereas I will use the most detailed subdivisions available: the neighbourhoods.

In this article I will use only a part of the same dataset of Kok et al: limited to workers who were mainly casual workers or dockworkers in Amsterdam. These are individual data gathered from the population register and poor relief archive. The people here represent the unskilled workers, one of the lowest income groups in the city. The names come from an unemployment relief agency of 1916 and a list of the harbour's employers union from 1920.[5] I have limited myself also to data that applies to the start and last address or that is constant.

For the city as a whole I have data on taxation and rent of neighbourhoods for several years. Some rent and tax data are only available on the highest level that I have termed districts. I will use this level for presentation of data. There is also a level in between called 'buurtcombinaties': combinations of neighbourhoods. The neighbourhoods were in time subdivided ever more until there were more than 200. They varied enormously in population though: from 300 to above 10.000. New districts were added in 1920 with the

2 See M. Wagenaar, *Amsterdam 1876-1914:* economisch herstel, ruimtelijke expansie en de veranderende ordening van het stedelijk grondgebruik. Amsterdam, 1990. chapter 7.

3 J. Kok, K. Mandemakers and H. Wals show that for the period 1910-1940 about 20% of the unskilled workers moved out of the old city. 'City Nomads: changing residence as a coping strategy, Amsterdam 1890-1940'. *Social Science History 29(2005)15-43.*

4 J.J Harts. en L. Hingstman. *Verhuizingen op een rij: een analyse van individuele verhuisgeschiedenissen*. Amsterdam, 1986.

5 So when I refer to the sample the data comes from the database on dockworkers using many sources.

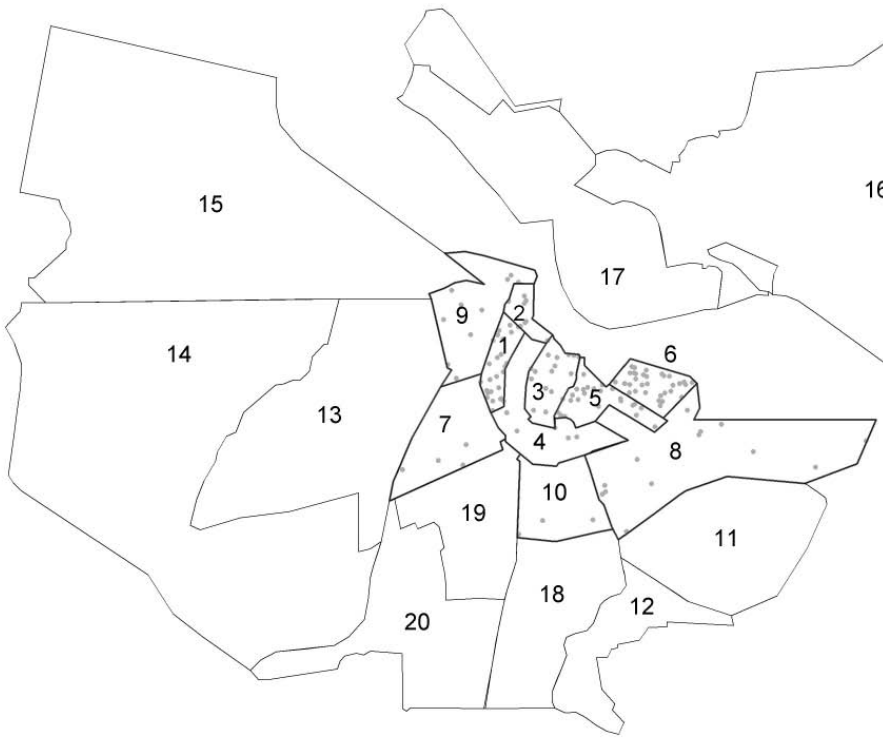
city's expansion until we end with 20 of them in 1940. Of the buurtcombinaties there were 50-56 from 1930 onwards.

In this paper I will just make a beginning and not use the rich set of individual variables that is in the database. I will limit myself to three questions.

1. What were the movement patterns in geographic terms? From where did they come, what was there destination?
2. Did the residential moves show changes in income as measured through the taxation scores of the neighbourhoods where they came from and went to?
3. Did the residential moves contribute to social segregation according to income?

Residential mobility

The general trend in Amsterdam after 1860 was for people to move out of the old city areas into newly built ones. The number of houses and people in the old city declined steadily because of renewal, slum clearing and enlargement of tenements. Of the sample the generation that married before 1909 was still heavily concentrated in the old city: only 21% lived in the new areas, mainly working-class districts west and east of the centre lying very close to the harbour. The Oostelijke Eilanden, Jordaan and Jodenbuurt were still the main districts were the dockworkers lived. These workers ended their lives in different places though (75% end in the 1930s): the new districts referred to (mainly IJ – Hugo de Grootgracht and Oost) had become their main living quarters. 63% had their final addresses in the new city. They had also spread to



Number	District
1	Jordaan
2	Westelijke Eilanden
3	Centrum
4	Grachtengordel
5	Jodenbuurt
6	Oostelijke Eilanden
7	Hugo de Grootgracht - Vondelpark
8	Amstel - Nieuwe Vaart
9	Spaarndammer- en Staatsliedenbuurt
10	Oude en Nieuwe Pijp
11	Watergraafsmeer
12	Oud- Ouder-Amstel
13	Stedelijk West
14	Landelijk West
15	Sloterdijk en IJpolder
16	Landelijk Noord
17	Stedelijk Noord
18	Amstellanen en omgeving
19	Vondelpark-Concertgebouwbuurt
20	Stadion-, Schinkelbuurt

Figure 1. Dockworkers in districts 1909 in Amsterdam

Table 1 First marriage cohort in the districts (figure 1)

	Marriages before 1909: first address	Marriages before 1909: last address	Zone	Class[6]
	%	%		
Centrum (3)	11,4	5,7	17thc.	Mixed
Grachtengordel (4)	2,9	0,7	17thc.	Elite
Jodenbuurt (5)	17,1	7,9	17thc.	Working-class
Jordaan (1)	18,6	5,0	17thc.	Working-class
Oostelijke Eilanden (6)	22,9	15,7	17thc.	Working-class
Westelijke Eilanden (2)	5,7	2,1	17thc.	Working-class
Oost (Amstel-Nieuwe Vaart) (8)	9,3	21,4	19thc.	Working-class
IJ – Hugo de Grootgracht (9)	7,1	16,4	19thc.	Working-class
Hugo de Grootgracht – Vondelpark (7)	2,9	2,1	19thc.	Lower-Middle-class
Oude en Nieuwe Pijp (10)	2,1	2,1	19thc.	Lower-Middle-class
Stedelijk gebied west (13)		3,6	20thc.	Lower-Middle-class
Stedelijk Noord (17)		9,3	20thc.	Working-class
Oud-Watergraafsmeer en Omval (11)		2,9	20thc.	Middle-class
Amstellanen en omgeving (18)		0,7	20thc.	Middle-class
Museum-, Concertgebouw-,Willemspark- en Apollobuurt (19)		1,4	20thc.	Elite
Landelijk Noord (16)		2,9	20thc.	Mixed
Total (N=140)	100	100		
In 19th & 20th c. districts	21,4	62,9		
In non-working-class districts	19,3	22,1		

Table 2 Tax score by cohort and zone of last address (%)

Marriages	before 1909		after 1908	
Tax score above mean	no	yes	no	yes
Zone last address				
17th c. inner centre	1,9	16,2	1,3	8,8
17th c. elite				2,9
17th c. working class	36,9	16,2	37,3	5,9
1860-1900	23,3	27,0	13,3	32,4
1900-1920	22,3	18,9	25,3	20,6
1920-1940	15,5	21,6	22,7	29,4
Total (%)	100	100	100	100
Absolute total	103	37	75	34
	%		%	
Start address	85,7	14,3	77,1	22,9
Final address	73,6	26,4	68,8	31,2
17th c. working class (start address)	96,7	3,3	95,7	4,3
17th c. working class (final address)	86,4	13,6	93,3	6,7
New districts (final address)	71,6	28,4	62,2	37,8

6 Categorized according to: working-class: >80% wage workers and <7% middle class tax payers; lower-middle: >70% wage workers and >7% middle class tax payers etc. in 1930.

Table 3 Taxation change from below mean taxation of movers in old city (%)

	Marriages before 1909		Marriages after 1908	
	Below mean	Above mean	Below mean	Above mean
stay in old city	79,5	20,5	81,8	18,2
to new city	66,1	33,9	63,2	36,8
Absolute total	72	28	39	13

the north part across the IJ, where new industries had sprung up. The cohort that married after 1908 started to live mainly in these new areas already (52,8%) and in the 1930s were spreading to the newest lay-outs of the city, leaving the 19th century districts: they went to the western and northern parts for instance. So, these poorest of workers were also following the general trend of spreading outwards to ever newer areas. And their movement outside the city centre had started in the 1910s. Still, of the second cohort there remained a sizeable fraction in the old city (32%). When we compare the neighbourhoods of the start and final addresses in our sample 14,1% ended up in the same area.

People of the first marriage cohort that ended in the old centre usually came from the other old districts. The Oostelijke Eilanden and the Jodenbuurt had the highest rates of stayers (25 and 29% respectively). Their inhabitants moved in the majority to the north-eastern areas. The Jordaan had very small numbers that stayed and the inhabitants went mainly to north-western harbour districts. Those that already lived in the new areas mainly stayed there. This east-west split continued in the movements of the second marriage cohort. Only the northern area was receiving people in equal amounts from both sides of the city.

Social and geographic mobility

To establish a score for income standing of the neighbourhoods I have computed the z-score of the number of people who were taxed in relation to the population of the area. These z-scores show how many standard deviations the percentage of people taxed was above or under the mean in a particular year. The number of people who were taxed increased in this period, so it would not do to record just the percentage of the taxed population. The z-scores can be used to compare different years with different means. I have used several years for which I have tax data: 1893, 1898, 1915, 1920, 1930 and 1936. In 1889 24% of the occupied population had an income tax assessment. In 1910 this was 46% and in 1930 this was 70%. The crisis of the thirties reduced this but I don't know how much.

Can the incomes at neighbourhood level be used to gain insight into the social standing and social mobility of individuals? Individual incomes in the sample are so diverse and incomplete that makes them hard to use. And occupation cannot give a quantitative indication of social mobility therefore I try to find out if the neighbourhood tax score can be used for this purpose. The main problem is that an income rate by neighbourhood washes away individual difference. But, as I will later show, the working-class neighbourhoods where most of the sample people lived were relatively homogeneous in their (low) income rates.

The tax scores of the neighbourhoods people lived in give a clear picture of the social standing of the addresses they lived at. The starting addresses of the married couples show that only 18,1% of them lived in a neighbourhood that had a tax score above the mean. The first addresses for the marriage cohort before 1909 are between 1890 and 1909. The final addresses of the whole sample that have been recorded before 1941 show an increase to 28,5% that have tax scores above the mean. Many addresses recorded an upward change. This being said these data show that these people still lived in the poorest of the neighbourhoods in the city.

Were the ones that stayed in the centre the poorest? Table 2 above shows that those who lived in the old city primarily did so in under average taxed neighbourhoods. In the new districts this was less the case. The people that moved to the new areas must have been rising in tax terms then? Table 3 shows that those that started in a taxation area below the mean in the old city were better off in the new districts than in the old ones. In terms of absolute change those that remained in the old city had the highest increases in the second cohort. This is because the highest rises in the number of people taxed between 1920 and 1930 can be found in the old city's working-class districts. But they remained the poorest districts.

Social segregation

Did social segregation increase? Let us compare the city data on the district level with the patterns of the sample.

A quick indication of segregation can be gained from a summary measure like the index of dissimilarity.[7] As this index increases with the number of categories (districts) it is of limited use for comparisons in time with changing categories. For a comparison of the distribution of city population and the sample at two points in time it will do. For the city population of 1900 and the start addresses of the first marriage cohort the I_D is 30, meaning that this amount (%) of the population was in different districts. The first marriage cohort is overrepresented in the old city and the workers districts. While 58,6% of the total population lived in the old city this was 78,6% for the sample. For working-class districts the ratio was 54% and 80,7% for the city and the sample. We have seen that the first marriage cohort spread out over the city. While in 1936 in the whole city the population in working-class districts and the old city was 50% and 19,4 respectively, for this cohort it was 77,9 and 37,1. For the second cohort the same values for the final addresses were 62,4 and 32,1 respectively. The I_D for the final addresses increased for the first cohort to 42, for the second cohort it came from 24 to 30,4. The comparison between the sample with many addresses and two points in time is an uneasy one, but it is clear that the dockworkers had their places not in the south far from work or in the elite districts. In the first cohort all the measurements point to increasing segregation, in the second the figures are contradictory. In social geographic terms the dockworkers of the first cohort clustered together in five broad areas of working-class type (table 1).

What about segregation according to income? Above and in table 2 I have already mentioned that the sample population was heavily concentrated in neighbourhoods with very low tax rates. They showed some upward movement in both cohorts that can be interpreted as decreasing income segregation, but this was probably due to a general rise in income. For the sample there is the indication that those who remained in the centre were among the poorest. Of the city as a whole one can speak of a tendency toward desegregation, at least after 1920, meaning that the general rise in income meant the spread of (lower class) tax payers across districts. This did not mean that the working-class districts were stepping up in the income hierarchy, but that they gained more tax payers.[8] Up to 1915 segregation had increased. In 1878 50% of the districts had more than 80% of low-class tax payers; in 1915 this rose to 75% of the districts; in 1930 the figure had fallen to 68% of the districts which had increased in number by then. The working-class districts where the majority of the sample population ended were rather homogenous in income division: they mainly had lower class taxpayers and a large class of non taxpayers (40-50% of the occupied population in 1930 whereas the richest area had only 5% non taxed!). In 1915 they did not have more than 2% of mid category tax payers. In 1930 some had risen to around 7%. City means in these years were 8,8% and 18,5% respectively. Their share of the elite tax payers was very small as you can see from figure 2. It also shows the lack of diversity in house rents in the poorest districts (Oostelijke Eilanden, Jordaan, and IJ-Hugo de Grootgracht). An area like Amstel-Nieuwe Vaart was not so homogenous, because it was a mixture of working-class and middle-class districts. I lack rent details to split it up.

The elite and middle categories were also concentrated in a few areas, but it spread out more after 1915. The newly built districts in the 1920s gained a large share (26%) of middle category tax payers by 1930. The main elite area switched from the centre (the canal zone Grachtengordel) to the south (Vondelpark-Concertgebouwbuurt). In 1915 the old city contained 41% of elite tax payers, but this fell to 18% in 1930. The new elite district Vondelpark-Concertgebouwbuurt contained 36% of elite tax payers in 1915 and increased its share to 41% in 1930. From figure 2 you can see that it was much more homogenous in its rent structure than the old elite district the Grachtengordel.

These income divisions were not between the old and new areas as such. No significant differences in income levels for the city as whole between the old city districts and the new districts or between the old and new working-class districts (measured at the level of

7 The formula is ID = 0,5 * $\sum$ |Xi-Yi|. Use percentages for the share of the entities in the total. Sum the absolute differences between the percentages of the two populations you want to compare and multiply the result by a half.

8 See my paper 'Social class and area differences in fertility decline in Amsterdam, 1850-1940' for all the data on tax income segregation. At: http://www.niwi.knaw.nl/home/henkl/

9 Rents from *Verslag over den toestand der gemeente Amsterdam*, 1914. p. 248. The formula for rent hoogeneity is: $\sqrt{\sum}$ **Xi2.**

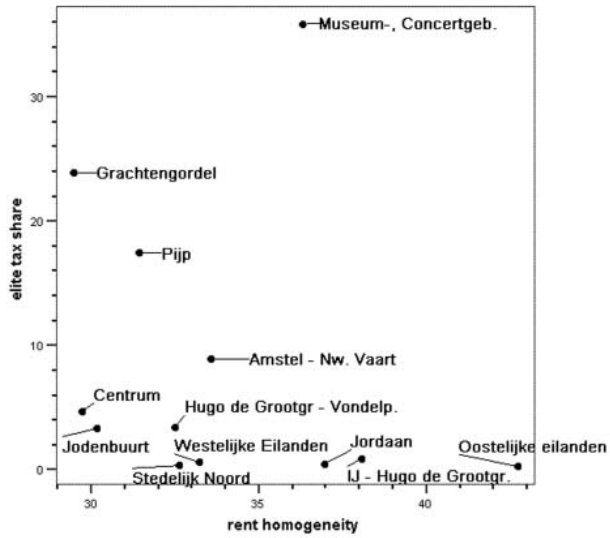


Figure 2. Elite taxation segregation and rent homogeneity 1915 (%)[9]

buurtcombinaties) could be established through *analysis of variance*. This means that variation in incomes inside the zones or inside the working-class districts was bigger than between them. If there was no significant difference in tax incomes between the working-class districts in the zones, there were perhaps differences in rents that mattered?

Housing and rents

A survey of the municipality in 1925 of all the housing in the subsidized sector (housing corporations, woningwetwoningen) amounting to 13.857 tenements shows 363 dockworkers and 717 casual workers living there.[10] They occupied mainly housing in the northern part of the city (across the IJ) and in the districts near the harbour like the people in the sample. Most of these workers were living in council houses, having the lowest rents and built for the poorest. They generally paid there a mean 16 and 18% of their income on rent. The houses built before the First World War had lower rents than those built during and after the War. People living in the older houses paid a lower percentage of their income on rent than in the newer houses. In the first a mean rent of 281,85 guilders per year was paid while in the second 347,57 guilders was paid in 1925. As the mean income of the inhabitants was 1.928 and 1.938 respectively, the mean percentage of income paid on rent in the new houses was higher: 15% compared to 18%. The mean rents in old city housing were lower than in the new city, and any distribution shows that the lowest rents were found there.[11] If unskilled workers in the old and new city had no real different incomes then they must have been spending more on rent in the new areas. Workers kept finding it difficult to move to these higher rent areas: they did so mainly because they had a young family that demanded better housing as an investigation from the 1930s concerning people in the lowest rent category points out. This investigation shows that many moved inside their own district because they could not pay higher rents elsewhere. Even though, 71% of 1814 movers did move to a tenement with a higher rent.[12]

In 1930 there were still marked differences in health and housing conditions between the districts. In terms of residential density the inner city centre came up very high on the list. A mass of people were still living there in overcrowded tenements and rooms. Even clearer do the unhealthy conditions of the old city as a whole show up when we look at infant mortality rates. Here all the 19th and 20th districts show lower rates than the old city districts. What is especially interesting is that these differences cut across the income divide. Of course, these area rates do also point to the mixed character of the old city and might work out differently individually. However limited the social mobility of the dockworkers in the sample was, they did move to areas that were healthier and had better housing conditions than those in the centre and in that respect they bettered themselves.

Conclusion

The spread of the sample population of casual and dockworkers across the city was quite different from the population as a whole. Their place of living seems mainly to have been influenced by the rent of housing and the proximity to work. When one looks at the total of residential moves then the impression is

10 De verhouding tusschen inkomen en huur in de vereenigings- en gemeentewoningen te Amsterdam. Gemeentelijke wonigdienst Amsterdam. 1925.
11 See: *Statistisch Jaarboek Gemeente Amsterdam*, 1940, table 10, p.48.
12 Waarheen zijn de bewoners der op 1 juli 1936 onbewoond gevonden woningen met lage huren verhuisd? Gemeentelijke woningdienst Amsterdam. 1937.

formed that these people remained locked in a very narrow area especially inside the old city. This was not the case: they too moved out of the old city like many others did. If one takes up the perspective to see the residential moves as elements of a living strategy then moving to the new parts of city with better living conditions was an illustration of it.

The increasing segregation in Amsterdam that developed between 1878 and 1915 halted somewhat after this date, but did not really change structurally. Segregation according to income and social class between the districts was evident. The workers in the sample did inhabit those areas on the lower end. The neighbourhood tax scores show quite clearly that the sample people lived mostly in the lowest income areas and as these working-class districts were relatively homogenous in their income division they show the segregation according to income. Those who remained in the old city were almost wholly concentrated in areas with taxation below the mean.

Using the neighbourhood tax scores as a measure of social standing was useful to gain insight in the social segregation of the individuals but as a measure of social mobility it has its limitations. We cannot adequately explain individual changes with data on this level. The next step will be to use all the individual variables to further explanation of the pattern of segregation. Still, the geographic mobility of the dockworkers and casual labourers to the new areas does point to a certain social mobility.

Historical geographic data dissemination through the web

*Rui Lopes**

The site Atlas and future developments towards its interoperability

The late 90s development in internet technology had made possible the integration of new information technologies on the web. Geographic Information Technologies (GIT) are among them and different software providers made available a variety of tools to embed Geographic Information (GI) in web sites. The SIGMA team recognized the added value of this new technology and seized the opportunity to develop the Atlas, Historical Cartography web-site (Atlas). This paper has the purpose of presenting the methodology used in the Atlas's web enabling process and to explore a possible route to take advantage of recent developments in GI interoperability to improve data usability.

1 Used methodology

The development of the *Atlas* had been guided by the use of hierarchic administrative boundaries as the main navigation tool to historical census data and evolution of administrative boundaries. The navigation technique that had been chosen was considered the most suitable to allow the users to access the geographic units of a specific level that forms a geographic unit of an upper level. That option was possible because Portuguese administrative areas have a full comprehensive coverage of the entire territory, i.e. any land portion is always part of one and only one administrative area, at every different administrative level. Although that fact is absolutely true in present administrative boundaries, in past times there were exceptions and inaccuracies that made the development work more complex to handle with those exceptions. Some of our approaches to deal with that are explained under the heading 'Scalability and flexibility'.

1.1 Data validation and optimization

The SIGMA project used the polygon as the data acquisition unit in all the datasets production processes, which was considered more suitable to administrative areas and also the best approach without using more expensive topology-based GIS packages. Although that have been clearly the correct option, it made necessary some additional steps to make the information suitable to distribute through the web in a full customized application. In order to run topological validation of all the geographical data and to use generalization algorithms to decrease the information weight, it was necessary to transform the polygon features in a polygon topology, comprising links – the unique poly-

*Department of History, Faculty of Social Sciences and Humanities, New University of Lisbon, Portugal,
Contact: rui_m_lopes@yahoo.com

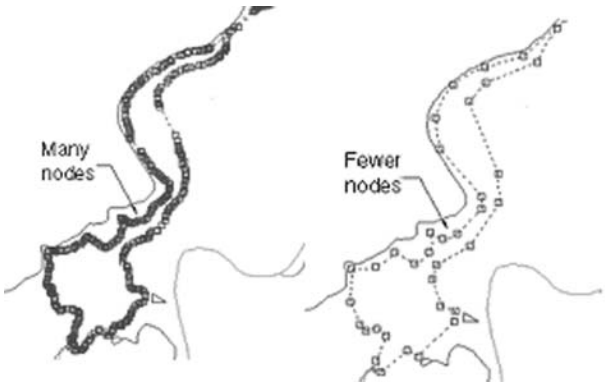


Figure 1. Example of the application of a generalization algorithm in line data.
Source: Autodesk Map Help system

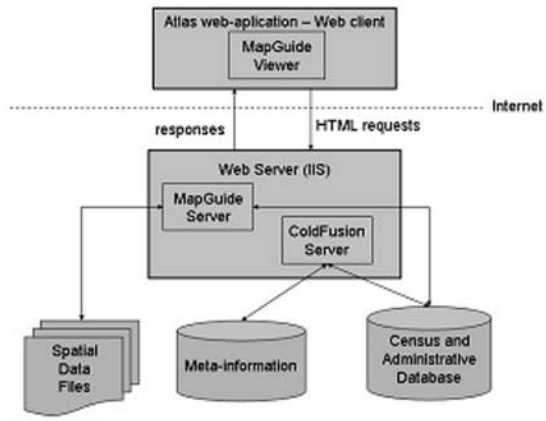


Figure 2. Atlas application model

gons boundaries – and centroides – the mass point of each polygon with its identification attached. After the topological verification, in which a few inaccuracies, otherwise imperceptible, have shown up, like small polygons overlaps and residual areas, the lines have been generalized to different levels. The SDF Loader algorithm[1] was used in different generalizations percentages and after analyzing the relation between the detail level and the information size decrease, it was considered that the initial limits could fall down to 20% of their initial granulity, i.e. its initial number of points per line. That represented, to the lower administrative level, a decrease from about 28Mb to 6Mb in the geographical dataset size.

After running the chosen generalization to the lower administrative level – called *freguesias* – the topology was rebuilt to find out and correct new errors made during the generalizing process. Only after having a full corrected topology, the polygons were recreated and used as the base to upper levels administrative boundaries. That process guaranteed that the different levels got 100% coherent boundaries, even after generalization. After automatically recalculating the area fields, the datasets were exported to Autodesk Spatial Data File, to web publish using Autodesk MapGuide.

1.2 Web site application model

The web mapping package chosen was Autodesk MapGuide based in performance and customizability criterias. Considering the high quantity of information to be published and the aim to develop advanced geographic data interaction tools, MapGuide was considered with its vectorial viewer as the most suitable solution. The application was developed using the web-mapping server in conjunction with ColdFusion Server as web-database server to access the census information stored in Access databases. Care was taken in defining indexes in all the searchable fields, defining coherent identifiers and avoiding information redundancy. The scheme below illustrates the application present architecture.

1.3 Scalability and flexibility

The SIGMA and Atlas projects have the goal to cover two hundred years of administrative change and demographic information. To aim to achieve that all at once would be unrealistic and hardly successful. The *Atlas* project was launched in 2001, comprising the administrative boundaries between the 1801 and 1855, and the census information for 1801 and 1849. That information was published to three different ad-

1 David H. Douglas and Thomas K. Peucker developed the algorithm used by Autodesk MapGuide SDF Loader at the University of Ottawa and Simon Fraser University, British Columbia. More information can be found at their paper 'Algorithms for the reduction of the number of points required to represent a digitized line or its caricature'.

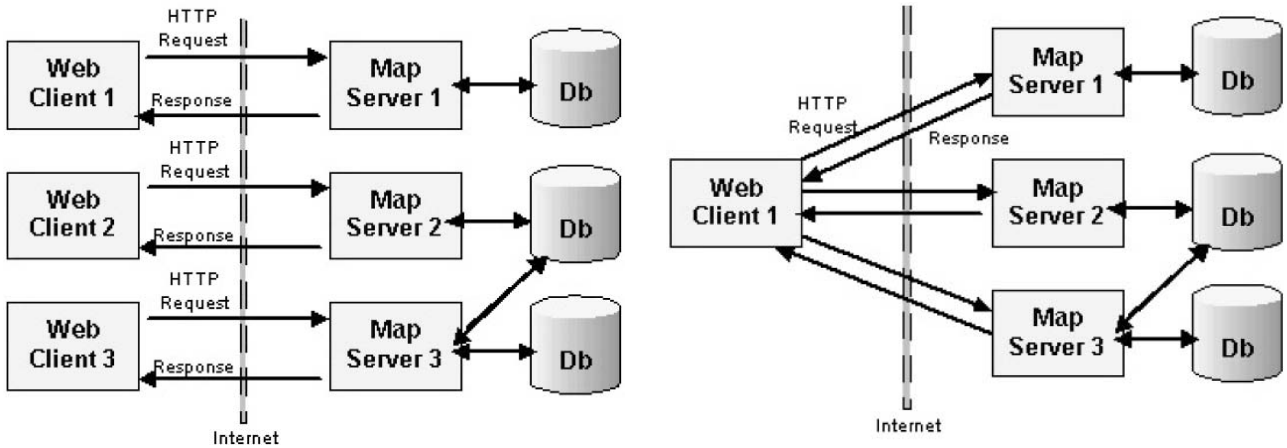


Figure 3. Comparison of interoperable and non-interoperables map servers. Source: OpenGIS® WebMapServer Cookbook.

ministrative levels – a lower level, called *Concelho*, a medium level, called *Comarca* or *Distrito* accordingly to the time period, and a upper level, called *Provincia*. Presently the *Atlas* already comprises 100 years of administrative boundaries, and includes a new lower level, called *Freguesia*. The importance of a scalable and flexible data model has been crucial to the project success. With limited changes in the application code, it was possible to integrate the new datasets, without compromising the previously defined hierarchic navigation. The application menus are full dynamic, being generated in HTML based on information stored in a meta-information database. When a record indicating the existence of a new administrative level is added in the meta-information database table and the corresponding datasets are published to a web-accessible folder, the new information would be immediately available to users.

On the other hand, information had not always been produced so carefully as nowadays and often exceptions and missing data are found. For example, although a *Comarca* is always an aggregation of *Concelhos*, in 1801 there was one *Concelho* that was part of two different *Comarcas*. It was necessary to create application code to handle with this exception in a coherent fashion without compromising the correctness of the information published. The importance of flexibility, at both the data model and web application levels, has been a key issue to the project maintenance and evolution.

2 Future development and interoperability

The *Atlas* project has a limited user community, mainly in academic environments formed by social sciences researchers and students. But the information itself could interest a higher public if it was possible to compare administrative areas with other geographical information, like rivers, altimetry, present population distribution, etc. The distribution of the *Atlas* datasets through the web is being analyzed as a possibility using the Open GIS® Web Map Server (WMS) implementation.

The Open GIS Consortium Inc (OGC) is *'...an international industry consortium of 278 companies, government agencies and universities participating in a consensus process to develop publicly available interface specifications. OpenGIS® Specifications support interoperable solutions that 'geo-enable' the Web, wireless and location-based services, and mainstream IT. The specifications empower technology developers to make complex spatial information and services accessible and useful with all kinds of applications.'*[2] The Geographic Markup Language (GML), a XML language to encode geographic information, has been developed and adopted within the OGC with the intent of *'providing a normative encoding for geographic information in XML that can be used for geo-spatial data interchange, and for web-based geo-spatial information networks' (OpenGIS® WebMapServer Cookbook)*. Although the high level of ion of the OGC specifications, the different software pro-

2 In OGC web-site

viders have been launching implementations of them. The *Atlas* is supported in a web-mapping package that already has an OGC Web Map Service compliant extension, making possible the dissemination of the *Atlas* datasets using the WMS technology. The next two schemes exemplify the potentialities of geographic information share using the WMS technology.

Through a group of requests – GetCapabilities, GetMap and GetFeatureInfo – it is possible to retrieve information from any WMS compliant map server using different map-server packages, without restrictions like scale, projection, earth coordinate system or data format.

From the different possible approaches already prototyped and tested to implement WMS data sharing, the development of a new web-client to retrieve Atlas maps to overlay with others maps would probably be the best one. Without getting into detail, *Atlas* map requests would have to include administrative level, time period and variable represented (if a census map is requested).

Due to the technological requirements of the project, it would be beneficial to the Atlas project team to define partnerships with other organizations. Two partnerships with organizations matching specific criterias would particularly suit the project goals: 1) an organization with a stable and robust web mapping application, with interest in sharing geographic data with the Atlas project; 2) an organization with good know-how in XML and web-technologies interested in developing a web client to access geographic datasets through WMS.

References

Harder, Christian. 1998. '*Serving Maps on the Internet – geographic information on the world wide web*', USA, California. ESRI Inc.

Kolodziej, Kris (ed), '*OpenGIS® Web Map Server Cookbook*, OGC™ Open Geospatial Consortium Inc. ENovember 4, 2004. Working paper, draft version.

Beaujardiere, Jeff de La (ed) *Web Map Server – OGC Standard*, OGC™ Open Geospatial Consortium Inc. August 2, 2004. Standard specification, final version.

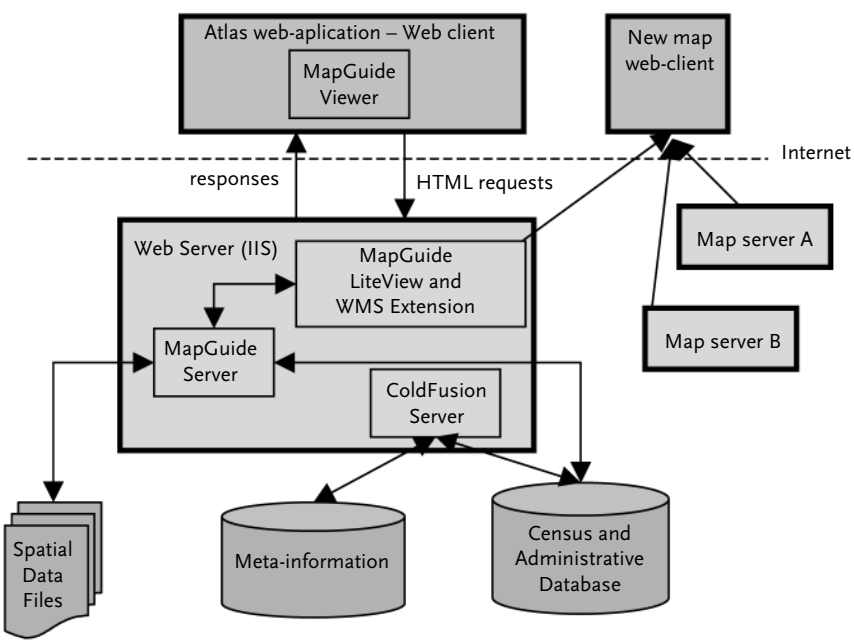


Figure 4. Atlas application model after data interoperability enabling

Reconstructing lost spaces. Affordably, that is

*Peter Melms**

1 What are lost spaces and why to reconstruct them?

Virtual Reality exhibits in museums are not exactly ubiquitious today, but there *do* exist examples of very impressive reconstructions, mostly of buildings, in blue rooms. One of the most impressive examples might be the temple of Zeus reconstructed at the Foundation of the Hellenic World in Athens[1]. Besides the educational and presentational value of such systems, there do exist studies which have tried to use 3D simulation techniques for analytic purposes. Mainly in the sense, that a building or a historical setting, which does not exist or does not exist in the form any more, in which it has been described in a historical document, is reconstructed virtually. Frequently such reconstructions have also been discussed as experimental settings in which various theses about a historical situation could be tested by arranging lighting, movable objects and similar components, successively in such a way, that a configuration is reached which explains the surviving description best.

While conceptually attractive, these techniques have by far not been used up to capacity, as the technical infrastructure needed is still much too expensive and for most such experiments dedicated projects would be needed. (Though CAD tools, like AutoCAD, which have been used quite frequently in art history and related subjects, cut the project costs very significantly, compared to full VR projects.)

Indeed, there could be a less expensive way of reconstructing the lost spaces mentioned above: The increasing capability of today's computer hardware allow advanced 3D simulation even on recent mainstream pc's.

Modern game engines make use of these skills to create virtual worlds with a rising degree of perfection. For instance, they offer integrated physics systems, allowing almost lifelike behaviour for displayed objects.

In principle, there is no reason, why these tools could not be used to create 3D simulation, specifically also in the analytic sense mentioned above, which are cheap enough to become a standard tool for Humanities research.

In the present case, a prototype of a low cost reconstruction of the painted cave of Altamira has been created, which allows to experiment with various types of lighting to understand better in which visual situation the cave has been at the time the paintings have been created. The next chapter gives a short overview on the available software and standards.

2 Available software and standards

There is a huge number of software packages and tools, used to create 3D-Content. In this chapter, with special regard to the requirements for the reconstruction of the Altamira cave mentioned above, some of them will be introduced to provide an overview to this broad and fast developing field that might be fragmentary for the same reason.

2.1 VRML and X3D

One of the best-known formats of VR-type software to be mentioned here is the VRML[2] format. The acronym VRML stands for *Virtual Reality Modeling Language*. VRML came up in 1994. It was designed to provide a cross-platform standard for interactive three-dimensional graphics across the World Wide Web.

It is this multi-platform approach that turns out to be the greatest benefit and one of the major drawbacks at the same time. On the one hand, any user is

*University of Cologne

1 For further information see: http://www.fhw.gr/ or http://www.fhw.gr/cosmos/en/vr/.

2 For further information on VRML see also: http://www.web3d.org/x3d/specifications/vrml/ISO-IEC-14772-IS-VRML97WithAmendment1/.

able to watch VRML-contents using his browser and a plug-in, no matter which operating system is running on his computer.

On the other hand, due to this approach, VRML underlies some great limitations. To be executable on almost any PC, there were constraints to be made. As VRML is a text- and scene-graph based language, the complete rendering takes place on the client-pc. Therefore, the ability of lighting, especially of dynamic lighting and shading is rather restricted. So it is not possible[3] to create moveable, dynamic light sources, which cast dynamic shadows.

X3D is the successor to VRML, which is still under development of the w3-consortium. Even if X3D provides some advanced features concerning graphical presentation, like bumpmapping and hardware acceleration, it shares the restrictions on the skills for lighting mentioned above.

2.2 Java 3D

Java 3D[4] is an Extension to the Java 2 Standard Edition that creates a connection between the Java Runtime Environment (JRE) and the computer's 3D graphics support. Like VRML, it is scene-graph-based and able to use either OpenGL or DirectX for low level rendering, to take advantage of 3D hardware acceleration. In principle, with Java3D as a high level programming API, it is possible to create an advanced 3D-Engine that meets all the demands for building an interactive, dynamic lighted model of the Altamira cave. The model of the cave could be created using a 3D authoring tool like 3ds Max and afterwards being imported into Java3D in use of the provided Loader Classes.

The major drawback here is, that, for a novice to programming, it takes up to two years or more, learning to develop a full featured 3D engine yourself. It is quite obvious, that this approach is not exactly applicable for research in an environment given at universities, for the process of programming, finding errors and debugging the code is very time consuming.

Hence, the question emerges, whether there is a more time cutting way of creating a 'lost space' like the Altamira cave, lighting it and making it accessible interactively.

2.3 Crytek CryEngine

Recent game-engines, supported by the increasing performance of the latest graphics processing units (GPU), have reached an unparalleled level of visual quality and realism. In search of a low cost and quicker way to create lost spaces in an adequate visual quality, sooner or later these game-engines come to the fore, providing quite effective algorithms for the creation of internal rooms.

The CryEngine[5] is a state of the art 3D-engine developed by Crytek a German interactive entertainment development company. It was first used in the game *Far Cry*. In order to keep the gaming community occupied, *Far Cry* is delivered together with a build in Level-Editor, which allows easy creation and integration of individual contents, offering 'what you see is what you play' feedback. Besides real time shadows, real time per-pixel lighting and an integrated physics system, it supports Shader Model 3.0, making full use of the hardware accelerated features of latest DirectX 9.0 graphics hardware.

Furthermore the CryEngine allows the usage of bump- and normal-maps, which enhance the visual appearance of textured objects exceedingly. 3D Objects could be created easily with popular 3D-authoring tools like *3ds* Max or Maya and being imported into the Editor using the provided exporter tools. Last but not least the Sandbox-Editor shortens development time to a great extent, making own programming work almost unnecessary.

3 The reconstruction of the Altamira cave

The palaeolithic cave of Altamira[6] is located in northern Spain, near the city of Santander. It was discovered in 1879 by Marcelino Sanz de Sautuola and is now one of the most spectacular sights of palaeolithic cave art. The paintings are dated to 13.000 – 14.000 B.C.

Some of the most impressive paintings are to be found in the so-called 'hall of the polychrome paint-

3 At least not without greater expenses.

4 For further information see the Java Sun oganisational web site http://java.sun.com/products/java-media/3D/java3d-features.html or http://www.j3d.org, a Java 3D Community Site.

5 For further information see: www.crytek.com.

6 For further information on the Altamira cave see the Museum of Altamira web site: http://museodealtamira.mcu.es/ingles/indexprova2.html

ed ceiling'. Amongst others, this painting includes a group of lying bison, which are applied on natural bumps on the surface of the ceiling, resulting in a very three-dimensional appearance.The objective was to build a three-dimensional model of the cave (more precisely a model including the entrance and the hall of the polychrome painted ceiling), texturizing it and importing it into a game-engine in order to enable interactivity. The subsequent tools were used for the reconstruction and simulation:

– Discreet 3ds Max 6[7] for building and texturizing the 3D model of the Altamira cave
– Adobe Photoshop[8] for the assembling of the ceiling texture.
– Crytek CryEngine and Sandbox-Editor for making the cave interactively accessible and lighting it dynamically.

3.1 Building the model with 3ds Max

Even if there already are replicas of the Altamira caves, one exhibited in the 'Deutsches Museum' in Munich, where only a part of the cave has been reconstructed and one build in the 'Museum of Altamira' in Spain, which covers the whole cave, the author was not able to receive more precise geographic data about the caves.

But, as far as this approach is to be considered exemplarily, that does not seem to be a big drawback, as it does not too much harm for our purpose, if the model is not as precise as it could be. So the decision was made, to fall back on an 'interpolated' model, based on a ground plan with plotted contour lines, which has been digitized by scanning it.

Afterwards the contour lines have been traced as paths in Photoshop and exported as Illustrator-paths (.ai-format). Then these Illustrator-paths have been imported into 3ds Max, where they were shifted in height, corresponding to the respective elevation marks on the ground plan.

At this time a rudimental terrain-model, consisting of lines in different levels, was achieved. Using the built-in terrain module of 3ds Max, the spaces between the lines were filled up with polygons, resulting in a three-dimensional wire-frame model.This terrain-model has been taken as a starting point for modelling the walls and ceiling, by appending additional polygons and meshes. Particular attention was paid on constructing the polychrome ceiling, which has been modelled separately. The mesh forming the ceiling is composed of circa 8000 polygons (which is about half of polygons forming the whole cave) in order to achieve a higher level of detail. After the wire-frame models of the cave and the ceiling were finished, the textures were to be applied.

3.2 Creating and applying the texture for the painted ceiling

The ceiling texture is composed of several photographs, originated from the 'Wendel collection'.[9] To create a contiguous image the photographs were arranged in Photoshop on the basis of a drawing of the whole ceiling. Subsequent the image was applied to the wire-frame model mentioned above. This wire-frame was adjusted in order to form the bumps the group of bison were painted on. The texture was polished by adding a bumpmap generating a three-dimensional impression and a relief-like surface.

The rest of the cave was textured with a simple rock texture, and an underlying bumpmap. Now the two wire-frames were connected and the cave as a whole object was exported to the native file-format of the CryEngine, using the provided Crytek Exporter for 3ds Max 6.

3.3 Integrating the cave into the CryEngine

The exported model of the cave could be imported to the Sandbox-Editor easily. Starting with a blank level we only see a water surface and a sky box, containing a global light source: the sun.

To be able to simulate and analyse various lighting situations, it was necessary to prevent the cave being lighted from outside, for the effects of the different lighting settings could be seen clearly.[10] After doing so, the inside of the cave is completely dark. Now you can start adding lights.

The engine offers a dynamic light entity, which could

7 For further information see the Discreet web site: http://www4.discreet.com/3dsmax/

8 See the Adobe web site: http://www.adobe.com/products/photoshop/main.html.

9 The *Wendel Collection* is an image archive of Ice Age cave art which was created by Heinrich Wendel. The collection is property of the *Neanderthal Museum*.

10 This could be done by applying a *VisArea*, which is a ready-made object delivered with the editor.

be adjusted in a variety of settings. Starting with the light colour and brightness, the so-called prefab light-shaders could be applied to create moving or flickering lights, like from a campfire or a torch. Using these dynamic lights, several lighting situations were created: one, where many small lights were scattered all over the floor for instance[11]. Another was simulating a large flickering campfire in the middle of the cave. To demonstrate the dynamic lighting and shadowing skills of the engine an experimental scene was created, providing a semi-moveable light source in the form of an hanging lamp, which swings due to the laws of physics when pushed aside.

Unfortunately, there is yet no possibility to create a fully moveable light source. But the Sandbox-Editor delivers a potential workaround to this. Using the 'What you see is what you play' ability mentioned above, a dynamic light could be positioned and adjusted, then switch to game-mode which is done pressing one key, switch back and readjust and/or reposition the light and switch to game-mode again and so on.

4 Conclusion

Using an exemplary 'lost space' the presented approach gives an example for a low budget project done in a time-cutting way, even though one does not have access to expert knowledge in 3D modeling or Virtual Reality, or even in 3D programming. The game-engine used here could be exchanged easily, if something better emerges. One advantage of the CryEngine is its ability of drawing large-scale outdoor environments providing a view distance reaching up to 2km. Even the integrated physics system offers great possibilities for simulation purposes. Altogether, the project presented here was finished within about 4 to 6 months including research on possible 3D engines, learning to handle 3ds Max and becoming familiar with the chosen Editor.

References

Adobe Photoshop

Adobe Web site: http://www.adobe.com/products/photoshop/main.html.

Altamira

Museum of Altamira: http://museodealtamira.mcu.es/ingles/indexprova2.html.

Beltrán¸ Antonio: *Altamira* ... Mit Aufnahmen von Pedro A. Saura Ramos. Hrsg. und mit einem Vorw. von Gerhard Bosinski. Sigmaringen: Thorbecke,1998.

Lorblanchet, Michel: *Höhlenmalerei: Ein Handbuch*. Speläothek 1, Sigmaringen, 1997.

Pastoors, Andreas und Weniger, Gerd-Christian: *Bilder im Dunkeln: Höhlenkunst der Eiszeit - Die Sammlung Wendel*. Hrsg. v. Andreas Pastoors und Gerd-Christian Weniger, mit einem Vorwort von Paul G. Bahn. Mettmann: Neanderthal Museum, 2004.

Crytek CryEngine

Crytek Web site: http://www.crytek.com

Crytek – CryEngine Featurelist: http://www.crytek.de/technology/index.php?sx=cryengine

Crytek, *CryEngine® Sandbox, Far Cry™ Edition, User Manual v1.1*. Coburg: Crytek GmbH, 2003.

CryMod – Far Cry Community Site: http://www.crymod.com.

Discreet 3ds Max

Discreet – Web site: http://www4.discreet.com/3dsmax/.

Immler, Christian: *3ds max 5: Produktivität, Realismus, Dynamik* / [Christian Immler]. Burkhard Müller. - München/Germany: Markt und Technik, 2003.

Winkler, Peter: *Jetzt lerne ich 3D-Design mit 3D Studio Max 3*. München: Markt&Technik Verlag, 2000.

11 Which is assumed to be the condition in which the cave paintings have been created.

Foundation of the Hellenic World
Foundation of the Hellenic World Homepage:
http://www.fhw.gr/ or http://www.fhw.gr/cosmos/en/vr/.
Java3D
Sun Developer Network Homepage – Java3D Featurelist:
http://java.sun.com/products/java-media/3D/java3d-features.html.
Java 3D Community Site: http://www.j3d.org.
VRML and X3D
Chris Marrin/Campbell, Bruce: 'Teach yourself VRML 2 in 21 days'. Indianapolis, Ind.: Sams Net, 1997.
Kate Fernie and Julian D. Richards, *Creating and using Virtual Reality – A Guide for the Arts and Humanities*. Arts and Humanities Data Service. Oxford: Oxbow Books, 2003.
Web3D Homepage - VRML: http://www.web3d.org/x3d/vrml/index.html.
Web3D Homepage – X3D Overview: http://www.web3d.org/x3d/overview.html.

Computational representation of semantics in historical documents

Vanesa Mirzaee[*]*, Lee Iverson*[*]*, Babak Hamidzadeh*[**]

Electronic versions of historical documents are limited in both the representations they allow as well as the level of computational manipulation they support. These documents usually have minimal advantages over their hard-copy counterparts: increased availability and facilitation of keyword search.

In this work, we present a novel approach using ontologies to represent the semantics within the knowledge found in historical documents. Our main purpose is to demonstrate the possibility of representing the knowledge found in these documents in a manner that allows further computational manipulation than that which is currently provided. It was of utmost importance for us to allow this information to be reused and shared amongst users as well. In order to do this we required a model where the representation of the concepts in our domain and their relations were captured and understood at a computational level.

To ground our work, we applied our proposed methodology to encode a historical document covering the evolution of the constitution of modern Iran [14]. Our implementation allowed us to get an overview of the general concepts in our case study and relationships amongst these concepts. Additionally, this model provided us with different methods for capturing and presenting dynamic and temporal aspects of the information. The model captures the changes that the relations undergo throughout time (dynamicity). The temporal aspect of the coded knowledge made our representation more accurate and realistic. Our methodology, implementation with reference to our case study, and the evaluation results are presented along with our proposed future work.

1 Introduction

The dawn of the digital information age has yielded a large amount of documents captured into an electronic form. However, these electronic documents have minimal advantages over their hard-copy printed counterparts [11]. Until recently, it has been assumed that the main advantage of digital documents over printed matter is the convenience of being able to find these materials without having to physically obtain them from a library or other repository [12]. However, we believe that this new technology has the potential to provide us with additional functionality than traditional printed matter does.

In the case of historical document archives, a wealth of historical information exists in digital form. However, these documents and the paradigms utilized to represent them are still limited in both the representation

*Dept. of ECE, University of British Columbia, Vancouver, BC
**Library of Congress, Washington DC

they allow as well as the level of computational manipulation they support. Once we have this information in a digital format, it is unclear as to how the user might interact with this information besides being able to print it and/or read it. It is reasonable to assume that a user wishes to obtain access to both implicit and explicit information contained in these documents. Instead of basing user queries on keywords, it would be ideal for digital historical archives to provide methods and techniques for posing and resolving historical queries and later grant access to the source of the claims used to resolve them. For example, a historian might want to query relationships between characters, institutions, events, and locations of these events. Significantly, it is vital to capture how these relationships change over time. In short, once we have these documents in a digital format, we would ideally like to be able to: (1) share and reuse this knowledge, (2) capture the semantics implicit in the documents, and (3) allow computational manipulation of the acquired knowledge. This would allow automatic reasoning beyond the simple queries and keyword search provided by current information retrieval methods.

In order to achieve these goals for historical document archives it is necessary to find a semantic foundation that is rich and powerful enough to capture and code both the implicit and explicit information within the documents.

A growing trend within the AI (Artificial Intelligence) research community toward capturing (coding) the semantic within the knowledge and creating knowledge models that can be shared and reused, has driven knowledge representation to a new era. The main approach for capturing and representing knowledge in this new framework is characterized by the use of ontologies. Ontologies are conventionally used for semantic-based information representation and retrieval [4]. They are often used to represent a domain by providing consensual agreement on the semantic of the knowledge that the domain is aimed to express. An ontology can be used to define rich relationships amongst different concepts within the domain. Additionally, it allows members of a community of interest to establish a joint terminology which will help achieve reusability and sharing of knowledge [29]. It is these characteristics inherent to ontologies that have led us to choose an ontological approach to code the knowledge within a historical document, thus allowing the user or a community to achieve the aforementioned goals. The ontological modeling presented allows sharing, reusing, and automatic reasoning by capturing the document's semantic content.

The remainder of this paper is organized as follows: Section 2 describes our motivations for using ontologies with reference to other currently available techniques for presenting historical knowledge. This is followed in section 3 by an overview of our methodology and implementation with reference to our case study. Finally we present our conclusions and proposed future work in section 4.

2 Why ontologies?

Narrative history texts contain certain characteristics that make them unique; they are usually organized along a time line and share similar key concepts such as people, places, events, etc. These documents all share one significant similarity: they are all about events that took place in the past [25]. Moreover, for any given period of time, these texts are not densely populated.

Until recently, most of the research related to digital historical document archives or digital documents in any other area of humanities, deals with how to create digital copies of these documents and store them in repositories (*e.g.* digital libraries) in a manner that facilitates later access to them [11, 12]. These digital archives usually integrate meta-data which provides information about their content or other attributes of the document [6]. The main functionality that these electronic documents using meta-data or other type of notation provide is the facility to retrieve the best-matched document to any search request [11, 12]. The field of information retrieval for this kind of document collections is an active area of research [1, 10].

Typically, these electronic collections are available in text or Hyper Text Markup Language (HTML) data formats exclusively. This is unfortunate since text and HTML data cannot be computationally manipulated in an effective manner.

It would be reasonable to assume that after obtaining a document a user might whish to obtain access to both implicit and explicit information found within its content. Obtaining this kind of information traditionally requires the user to read the document or at least some part of it. It would be ideal for digital historical archives to provide methods and techniques for pos-

ing and resolving historical queries and then providing access to the source of the claim used to resolve them.

The common approach utilized to represent the knowledge within a historical document is to use markup languages such as Standard Generalized Markup Language (SGML) and Extensible Markup Language (XML) which tag the information within the document [19, 25].

MEP (Model Edition Partnership) and HEML (Historical Event Markup and Linking) are examples of this general type of approach. In MEP, a series of models were developed to transform printed historical documents to electronic versions. This project employed SGML to support access to the information that exists within the documents. This was done by developing a set of SGML Document Type Definitions (DTDs) which define a markup system for publishing historical documents [19].

In HEML a markup scheme was used for digital representation of historical events on the web. In this work, XML tags were defined to describe historical events. HEML defines a lightweight XML language that associates web resources with historical events described in terms of time, location, and participants. The goal in developing HEML was to use this scheme as a framework to tag documents that record historical events, their date and location. This would facilitate assembling a computerized collection of this information and associating it to the document. It would then be possible to search for a description of events on a certain day, or in a certain region amongst collections that share this same scheme [25].

The current preferred approach for creating information infrastructures such as those presented in the last two projects is to use markup languages. Amongst these markup languages XML is the most accepted choice. However, XML can only represent the syntax within a document, not the semantic. This means that the tags in an XML document are semantically free. These tags do not have any predefined meaning; they do not represent the semantic, meaning, or behavior of concepts in the document and only represent content and structure [13].

These projects use XML as an exchange format. However, an XML document by itself is not completely useful for our purposes. Associating semantics to these tags requires additional mechanisms to describe the behavior and the meaning for the concepts within a document. This association can not be done unless there is a consensus amongst people within a community interested in a specific domain on what these semantics are [29]. In addition to this, in order to capture the dynamicity of the information in a historical document we require additional functionality to that which is attainable by using markup languages.

An ontology is the basic building block that can be used to define richer relationships between different concepts. It allows members of a community of interest to establish a joint terminology which will enable greater flexibility and help achieve reusability and sharing of the knowledge. We believe that an ontological approach would best fit our purposes.

3 Methodology and implementation

Building a well-developed, usable, and sharable ontology represents a significant challenge. There is great diversity in the way ontologies are designed as well as in the way they try to represent the world.

A range of methods and techniques have been reported in the literature regarding ontology building methodologies. However, there is still debate within the ontology community about the best method to build them [3, 17, 21]. Given that the knowledge to be captured is usually critically dependant on a combination of the domain and the applications being designed to exploit this knowledge [22], it is no surprise that these methodologies are primarily inspired by enterprise modeling or software engineering [8, 15, 16, 30, 31]. For our purposes, it was important to scale these methodologies down and adapt them to facilitate document coding.

We divided the ontology building process into the following stages:

1. Identifying the purpose, scope, and users
2. Domain analysis and knowledge acquisition
3. Building a conceptual (informal) ontology model
4. Formalization
5. Evaluation

With our method, we focus on an evolving prototype of the ontology. At every stage in this model, it is possible to go back to any previous stage of the development process, in order to satisfy emerging requirements. Figure 1 illustrates how these steps are related, and in what order they can be performed to complete the ontology building process. We made every effort to maintain the following criteria for each and every stage of the development process: Clarity; Coherence;

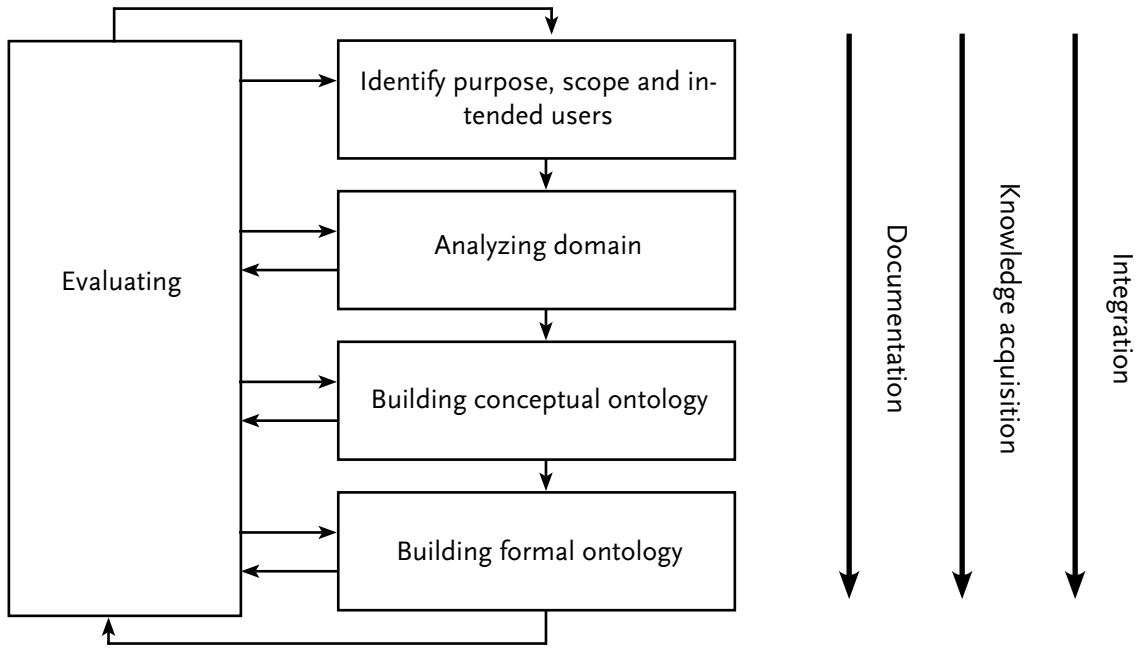


Figure 1. Our ontology development process. Integration, Knowledge acquisition, and Documentation are carried out throughout the entire development process.

Extensibility; Minimal encoding bias; and Minimal ontological commitment [9, 28].

3.1 Identify purpose, scope, and intended users

The main purpose for building this ontology is to capture the semantics of a historical document, in particular the temporal and dynamic aspects of the concepts and their interrelations. To promote sharing, reuse and enable better integration with existing knowledge sources we relied heavily on the consensual terminology available in general ontologies. The selected audience included the general public, historians and biographers who might directly access the semantic models.

The requirements gathered were formulated as a set of competency questions and motivating scenarios that our model must answer and provide support for. A few of these competency questions are presented in Table 1.

3.2 Domain analysis and knowledge acquisition

Using the competency questions and scenarios we then produce a set of concepts and terms covering the full range of information that the ontology must characterize to satisfy the requirements identified previously. In this phase, we use knowledge acquisition techniques such as brainstorming, in conjunction with informal analysis of the text to gather all potential relevant terms into a glossary [5].

This glossary includes the terms, their definition or description, and may include additional information, such as examples that help understanding these definitions. In order to provide definitions for the terms, we consulted dictionaries such as the Merriam Webster Dictionary and the Oxford Dictionary as well as general purpose ontologies such as SUMO [27] and WordNet [32]. Table 2 shows a partial view of our glossary of terms.

3.3 Building an informal ontological model

Once we have a relatively complete glossary of terms, we identify concepts, relations within the concepts, and their attributes. We use the guideline provided in [22] to do so. The results are stored in document tables called the Concept Dictionaries [5]. At this stage, the concepts are structured into naturally occurring groups using a combination of the approaches introduced in [22] and [16]. We categorized our concepts into five concept dictionaries relating to people, places, events, documents, and time. Each of these categories

Table 1 Some Competency Questions

1. Who was *Person* P?
2. In What *Events* was *Person* P involved?
3. What *Positions* did *Person* P hold?
4. When did *Person* P held these *Positions*?
5. Who was taking over *Person* P's *Position* Po?
6. What was the governmental position hierarchy at the time *Person* P holds *Position* Po?

Table 2 Partial view of glossary of terms

Term	Definition	Resource
Person	An individual, someone, somebody. An agent with certain rights ans responsibilities and the ability to reason, make plans, etc. This is essentially the legal/ethical notion of a peron	WordNet & SUMO
Government	A ruling body of a country	Oxford Dictionary
Position	A formal position of responsibility within an organization	SUMO & Merriam Webster
Governmental Position	A formal position of responsibility within a governmental organization	SUMO

holds the concepts that are most related.

For the next step, we use the previously generated concept dictionaries, along with the motivating scenarios and a middle-out approach to develop our graphical conceptual ontology model. Our conceptual model not only represents the concept taxonomy but also the other (non-taxonomic) relations that hold amongst the concepts within our domain.

Throughout the ontology building stages, we queried existing ontology libraries, such as Ontolingua [23], DAML [7], and SUMO [27] to search for similar or related terms and relations that might be useful. This was done in order to speed up the development process as well as to gain a better insight of how to build a particular area or set of concepts within our ontology. Thus we were able to build our ontology on a well-grounded structure. In particular, the time concepts were derived from general time ontologies [23, 27, 33] and the temporal relations in TELOS [20]. Events were based on Sowa's thematic roles or case relations [26]. Places were defined using standard ontologies for geographic information representation and categorization [2,18].

Figure 2 shows the top-level concept hierarchy in our domain. We identified five central concepts within our ontology: AGENT, PLACE, EVENT, DOCUMENT, and TIME. Every other concept in this domain is defined around these primitive concepts.

An important characteristic of the proposed ontological model is its ability to represent temporally dynamic concepts. This is of particular importance for historical data since the concepts and the relations between them change and evolve through time. This is accomplished by associating a time interval with each relation, as was done in Telos [20].

The representation of time is fundamental to any knowledge model that aims to represent change or action. Many ontologies that try to represent time are currently available. Some examples of these include: Simple-Time Ontology form Ontolingua Sever [23], Time Ontology from SUMO [27], and the Time ontology developed in the Stanford Knowledge Systems Laboratory [33]. In order to develop our time model, we studied these existing time ontologies. We based our work on these ontologies and adapted them to fit our needs. Traditionally time ontologies have a resolution of seconds or even milliseconds; however our time model has a coarser granularity and measures time by day.

Additionally, this model not only captures the rela-

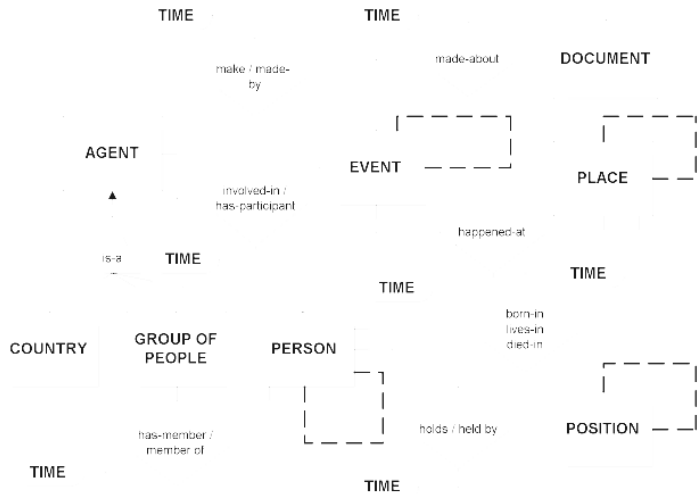


Figure 2. Overview of main concepts and relations in our history ontology. Dotted-lines denote the existence of type reflexive relations within a concept. The time tag on a relation indicates that a particular relation is time-dependent.

tionships between the concepts but also demonstrates the interrelated hierarchal structure within them. An example of such hierarchical structures found within our document is the governmental position hierarchy. In this hierarchy, not only do the people that hold positions change but the structure itself evolves throughout time.

3.4 Building a formal ontological model

The next step in our approach was to build a formal ontology based on the conceptual model. After a review of available ontology development environments, we selected Protégé-2000 [24] to formalize and instantiate our ontology. Our selection was based on the tool's expressiveness, flexibility, customizability, scalability, extensibility, and usability. Significantly, it also provided us with the facilities to test and evaluate our model.

Additionally, Protégé-2000 provides facilities to impose constraints to concepts and relations. While creating the ontology, it is necessary to make general assertions about fundamental concepts, and be able to later test and ensure these assertions hold across the entire knowledge-base. For example, in our ontology it was useful to assert common-sense constraints such as:

– All instances of Person have exactly one birth-date.
– A Person's birth-date must precede the death-date.
– Every Event in which a Person is involved, must take place between his or her birth-date and death-date.
– For any given time interval there can only be one person holding the position 'king'.

Figure 3 illustrates part of the governmental hierarchy that holds for a given time interval.

3.5 Evaluation

After designing, building, and formalizing our ontology using Protégé and enforcing constraints on attributes and relations, we used the knowledge acquisition forms provided in Protégé to instantiate our history ontology. Over seven hundred and fifty (750) instances were extracted from the history book and included in our ontology. Amongst these instances we find people, places, documents, and events.

In order to evaluate the correctness and completeness of the created ontology, we use the query and visualization facilities provided by Protégé-2000. We use the built-in query engine for the simple query searches and use additional query plug-in provided to create more sophisticated searches. We also use protégé-visualization plug-ins to browse the ontology and ensure its consistency. Visualization aids were particularly helpful when trying to understand hierarchical relations.

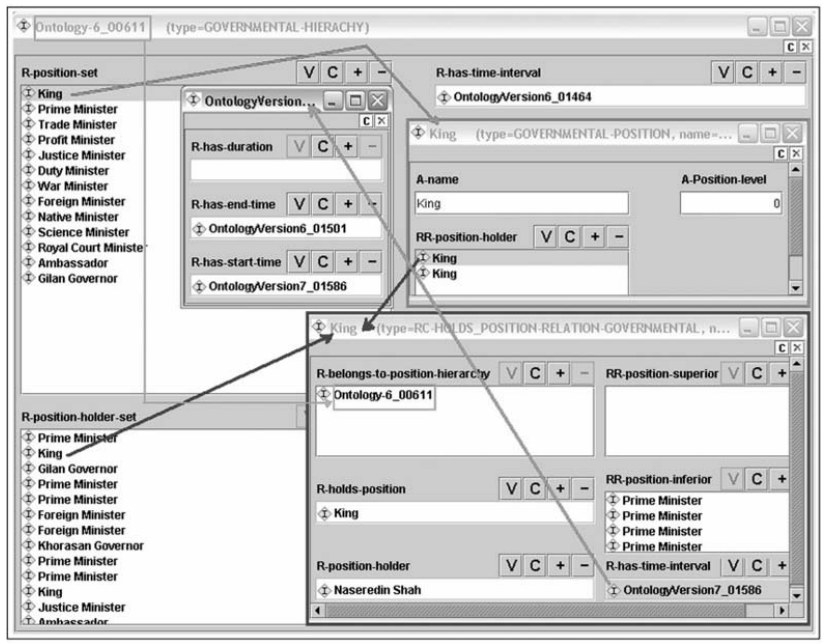


Figure 3. This figure illustrates an instance of the position hierarchy that holds for a specific time interval in our domain. In this example, the position 'king' (top right window) was held by a person called 'Naseredin Shah' (bottom right window) at that time interval. We can also see the inferiors and superiors of that person in that specific time period (within the bottom right window).

4 Conclusion and future work

In this work we confronted the limitations of traditional electronic documents. In particular we were interested in capturing the semantics of a historical document to allow for richer retrieval, reuse and manipulation of its embedded knowledge than is capable with standard text manipulation tools.

After adapting existing methodologies to the problem of text coding, we developed an ontology motivated by historical and biographical needs and the contents of the book 'The History of the Iranian Constitution.' Using this ontology, we then encoded the book's claims and verified our coding and ontology by proving that the ontology allowed us to answer all of our category questions. Our implementation allowed us to get an overview of the general concepts in this book, relationships amongst these concepts and provided us with different methods for visualizing dynamic hierarchical structures of both governmental positions and geopolitical interdependencies. Additionally, this model captures the changes that these relations undergo through time (dynamicity). The temporal aspects of the knowledge we captured proved to be useful in making our representation more accurate and realistic.

In order to facilitate the utilization of models such as the one developed here, we will require applications that facilitate interacting with this information. One challenge will be to develop easy, intuitive interfaces to both access and query these models that will allow both sophisticated and naïve users to take advantage of the information they encode. In addition, we hope to develop ontology development tools that reflect the methodology developed and facilitate its application to new domains.

References

[1] AHDS 'Arts and Humanities Data Services' http://ahds.ac.uk/.

[2] H. Alani, C. Jones, D. Tudhope (2000). «Ontology-Driven Geographical Information Retrieval.» *GIScience2000*, 2000.

[3] H. Beck, and H.S. Pinto, 'Overview of Approach, Methodologies, Standards, and Tools for Ontologies', The Agricultural Ontology Service (UN FAO), 2003.

[4] T. Berners-Lee, J. Hendler, and O. Lassila, 'The Semantic Web', *Scientific American*, May 2001, 284(5):34-43.

[5] M. Blazquez, M.F. Lopez, A.G. Perez, and N. Juristo, 'Building Ontologies at the Knowledge Level Using the Ontology Design Environment', *In Proceeding of KNAW'98*, Banff, Canada, 1998.
[6] J. Burchardt, 'Archiving the Internet - how to collect historical sources for the future.' *International Conference of the Association for History and Computing*, Poznan, Poland, 2001.
[7] DAML (DARPA Agent Markup Language) Ontology Library http://www.daml.org/ontologies/.
[8] M. Gruninger, M.S. Fox, 'Methodology for the Design and Evaluation of Ontologies.' *Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95 (Montreal), 1995.*
[9] T.R. Gruber, 'Toward principles for the design of ontologies used for knowledge sharing', *International Journal of Human-Computer Studies*, 1995, 43(5-6), pp. 907-928.
[10]HDS 'History Data Service' http://hds.essex.ac.uk/
[11] S .Hockey, 'Making Technology Work for Scholarship: Investing in the Data' *Technology and Scholarly Communication*, 17-36, 1999.
[12] S. Hockey, 'Electronic texts in the humanities: principles and practice' *Oxford University Press*, New York, 2000.
[13] M. Jarrar, 'Methodologies for extracting ontologies from the web.' *STAR Lab*. Brussels, Vrije University, 2000.
[14] A. Kasravi, *History of the Iranian Constitutional Revolution*, Amir Kabir Publicatoins, Tehran, 1940.
[15] M. F. Lopez, A.G. Perez, and N. Juristo, 'METHONTOLOGY: From Ontological Art Towards Ontological Engineering.' *Workshop on Ontological Engineering. Spring Symposium Series*: Stanford, USA, 1997.
[16] M. F. Lopez, A.G. Perez, J.P. Sierra, and A. P. Sierra, 'Building a Chemical Ontology Using Methontology and the Ontology Design Environment.' *IEEE Intelligent Systems & Their Applications*, 14(1), 37-46, 1999.
[17] M.F. Lopez, A.G. Perez, 'Overview and Analysis of Methodologies for Building Ontologies', Knowledge Engineering Review, 2002, 17(2), pp. 129-156.
[18] D.M. Mark, A. Skupin, B. Smith, 'Features, Objects, and other Things: Ontological Distinctions in the Geographic Domain', *Conference On Spatial Information Theory (COSIT)*, Morro Bay, CA, USA, 2001.
[19] MEP (Model Editions Partnership), Historical Edition in the Digital Age. University of South California. http://mep.cla.sc.edu/mepinfo/mep-info.html, 2003.
[20] J. Mylopoulos, A. Borgido, M. Jarrke, and M. Koubarakis, 'Telos: Representing Knowldege About Information Systems', *ACM TOIS*, 1990, pp. 325-362.
[21] N.F. Noy, C.D. Hafner, 'The state of the art in ontology design - A survey and comparative review', AI Magazine, 1997, 18(3), pp. 53-74.
[22] N.F. Noy, D.L. McGuinness, (2001). 'Ontology Development 101: a Guide to Creating Your First Ontology', Stanford, CA, Stanford University, 2001.
[23] Ontolingua, Knowledge System Laboratory, Stanford University, www.ksl.stanford.edu/software/ontolingua/
[24] Protege-2000, http://protege.stanford.edu/index.html.
[25] B. Robertson, 'The historical event markup and linking project (HEML)', 2003. http://heml.mta.ca/heml-cocoon/.
[26] J.F. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundation*, Brooks Cole Publishing Co, Pacific Grove, CA, 2000.
[27] SUMO (Suggested Upper Merged Ontology) http://ontology.teknowledge.com/.
[28] B. Swartout, R. Patil, K. Knight, and T. Russ, «Toward Distributed Use of Larg-Scale Ontologies», *Proceedings of KAW'96*, Banff, Canada, 1996
[29] P. Spyns, R. Meersman, and M. Jarrar, 'Data modelling versus Ontology engineering', *SIGMOD Record Special Issue on* Semantic Web*, Database Management and Information Systems*, 31(4), 2002.
[30] M. Uschold, M. King, 'Towards Methodology for Building Ontologies.' *Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95*, 1995.
[31] M. Uschold, M. Gruninger, 'Ontologies: Principles, methods and applications.' *Knowledge Engineering Review*, 1996, 11(2), 93-136.
[32] WordNet http://www.cogsci.princeton.edu/~wn/.
[33] Q. Zhou, R. Fikes, 'A Reusable Time Ontology', *Journal Proceeding of the Ontologies for the* Semantic Web *Workshop, AAAI National Conference*, 2002.

History ontology building: The technical view

*Gábor Nagypál**

The technical development of ontology based systems has been a topic of interest for computing scientists since the early 1990s. Over the years a number different ontology formalisms, methodologies and their supporting tools have emerged to help subject specialists build ontologies. However when attempting to create an ontology of a wide ranging subject such as history a number of specific challenges related to the nature of the domain arise. These include time dependence, subjectivity and uncertainty. This paper illustrates the challenges from the ontological engineers's point of view which were encountered in the development of an ontology of European history during the recently completed VICODI project. Solutions to those challenges are suggested based on our experiences during this project. These include the application of a fuzzy sets based temporal model and the introduction of intermediate models above the ontology implementation level.

1 Introduction

Ontologies play an important role in today's knowledge based systems. The Eurohistory.net portal, which was created during the VICODI project[1] is no exception in this regard. This portal provides a repository of resources about European history, and builds on the novel idea of visual contextualization. To put it in simple words, the idea of visual contextualization is the following: the context of historical resources (textual documents, pictures, videos etc.) which describes the relevant things to the actual resource in the domain of discourse, is represented in a suitable way. This resource context is visualized later to help users comprehend the meaning of the actual resource. It can be also used to raise the precision of searching for resources relevant to the actual one, which improves the user's navigation experience.

An ontology of European history plays a central role in this system. It sets the boundary of the knowledge space used to represent context definitions on the one hand, and encodes the required background knowledge to support different context based reasoning tasks, on the other hand.

Visual contextualization requires several nontrivial features (like automatic generation of context, finding relevant contexts etc.) which were implemented during the VICODI project. More information about those features can be found in (Surányi et al., 2004). The goal of the VICODI ontology was to support the process of automatic context generation and context matching, i.e. the ontology should be an applicationlevel ontology (Guarino, 1998). This means that the ontology had to store knowledge about European history in a form which is usable by intelligent algorithms. Support for human browsing of the ontology was not a priority for us. The purpose of this paper is to report on lessons learned during developing a fairly complex, application-level ontology of European history in the framework of

*FZI Research Center for Information Technologies at the University of Karlsruhe HaidundNeuStr. 10–14 D76131 Karlsruhe, Germany nagypal@fzi.de

1 The project was partially funded by the EU (EUIST200137534). Its home page is accessible at http://www.vicodi.org

the VICODI project. We will concentrate in this paper on the challenges from the ontological engineer point of view, the user or domain expert point of view of the process is described in detail in (Deswarte and Oosthoek, 2005a) and (Deswarte and Oosthoek, 2005b).

The structure of the paper is as follows. Section 2 represents the requirements to our ontology, and Section 3 discusses the special properties of history as a knowledge domain. Based on this Section 4 and Section 5 discuss the most important modelling and methodological challenges we encountered during the project from the ontological engineer's point of view. Section 6 concludes the paper.

2 Ontology requirements

Before starting the ontology development the goals of the ontology were clearly identified. As already mentioned, the main goal was to store historical background knowledge in a form which is optimally exploitable by intelligent algorithms for the purposes of contextualization and contextbased search. From that general requirement the following list of specific requirements were generated:

Simple, stable upper structure. It is clear that the more complicated the structure of an ontology is, the more complex algorithmic heuristics (or logical axioms) are needed to exploit it. Moreover, above a specific complexity level specifying such algorithms or axioms is not feasible any more. As most ontology formalisms are capable of inferring the most general concept of an instance, and the most general property of a property instance, it is not necessary to require that the ontology itself should be simple. It is necessary, however, that the ontology should have a stable, relatively simple upper level structure which algorithms can reason on.

Huge amount of historical facts. Algorithms need a huge amount of historical facts both for context generation and context matching. These facts can appear in the ontology as concepts, instances with their attributes, but most importantly as relations between instances[2]. In such a way Caesar can be related with Brutus, Lenin with the Russian Revolution and Napoleon with France. An intelligent algorithm can exploit these connections later on to infer that 'the emperor' refers to Caesar in a text about the Roman Empire, and refers to Napoleon in a text about the French Revolution.

General coverage of history. The ontology should cover history in general, and not only specific aspects such as the genealogy of the Habsburg Dynasty or the French Revolution. This requirement is very hard to fulfil, as history is as complicated as the culture of humankind itself. History contains science, philosophy, geography, potentially everything that was created in the past.

Please note that the first requirement contradicts the second and the third ones. It is clear that the more historical facts we encode and the broader coverage we have, the more complicated structure is needed to accommodate this knowledge in a meaningful way (at least if we want to present this knowledge with some precision). We believe that this dilemma, this tradeoff between complexity and expressiveness, is typical for any applicationlevel ontology which tries to cover a broad area of knowledge. In the VICODI project we set the historical scope of the ontology between 500 CE and today, that is we store knowledge about European history from early Medieval to the present day.

3 History as ontology domain

In addition to the already mentioned inherent conflict between fullcoverage of a complex domain and a simple, understandable ontology structure, history as a knowledge domain has many additional properties which are problematic from an ontological modelling point of view.

Time dependence. In history practically every fact is time dependent. For example while Strasbourg is part of France today, there were periods in the past when it was part of Germany.

Uncertainty. Because history deals with facts which are based on missing or contradictory historical documents, uncertainty is inherent in this domain. Primarily the time dimension is problematic, i.e. most of the temporal specifications are uncertain. As an example consider Stalin's birth date. O cially for the USSR it was 21 December 1879 but according to church records his birth was registered as 6 December 1878. Historians are in disagreement over which time specification is the correct one.

2 We consider also instances and instance relations as part of the ontology. In many cases the word 'ontology' is used only to denote concepts and properties, while instances and relations form the so called 'knowledge base'. The VICODI ontology includes both of them.

Subjectivity. Most complex historical notions are vaguely defined, and thus can be interpreted subjectively. For example notions like 'Russian Revolution', 'French Revolution', 'Middle Ages' or 'Industrial Revolution' have no clear start and end dates. It is even possible that there are periods which can be considered as only partially relevant to a complex event.

Why questions. Most knowledge representation formalisms are good at representing precise facts (axioms, rules) and ontologies are no exception, either. In history, however, those 'where, when, who, what'-types of facts are not really interesting – they can be easily looked up from encyclopaedias. Historians (and students studying history) are interested instead in 'why' types of questions. A typical historical question would be: 'How and why did the cultural image of the Jews change in medieval Europe?' However, as the goal of the ontology was to support the contextualization process, and not to provide an ontology which is interesting for historians (or for students) per se, we did not try to encode the answers to why questions into the ontology. Even if we tried to do it, it would be impossible, as there are not clear answers to those questions which can be captured using logical formalisms. In our case the whole VICODI system is capable to answer those questions implicitly by providing links to relevant documents containing partial answers for a complex 'why question'. Therefore this aspect will be ignored in this paper from now on.

4 Modelling challenges

The issues why a new ontology was needed on the first place, and what kind of strategy was followed during the ontology development process are discussed in detail in (Nagyp´al, 2004), (Deswarte and Oosthoek, 2005a) and (Deswarte and Oosthoek, 2005b). We discuss here only the most interesting modelling problems we encountered during the project.

History as a complex domain with unique features (see Section 3) caused some problems from the ontological engineering point of view, because it was not always easy or intuitive to encode historical knowledge using the relatively simple modelling constructs of present ontology languages. We implemented the ontology using the opensource KAON framework (Motik et al., 2002) which provides an extension of the W3C RDFS standard (Brickley and Guha, 2004). In particular KAON provides multilingual features – it defines many languagespecific lexicons around a language independent ontology core. There are also inverse, transitive and symmetric properties in addition to standard RDFS features which include defining a hierarchy of concepts and properties and specifying instances and their relations.

From the history specific challenges, time dependency alone did not cause difficult problems. Existence time of instances can be represented in a straightforward manner by connecting them to instances of a Time Interval concept (subconcept of a more generic Time concept). Time dependent relations can be represented using the standard technique of relation reification, i.e. by representing relations as instances of a TimedRelation concept (see Fig. 3).

The only issue which required a somewhat more advanced solution was the issue of the 'time dependent instantiates' relationship. For example it would be straightforward to have subconcepts of persons like King or President and say that Henry VIII was a king and Bill Clinton a president. It is not clear, however, how to represent the time dependency of those relations, as it is obvious that a person is not a king or a president for their whole life, at least usually not. We solved this problem by introducing the Role concept, and modelling the particular roles of individuals[3] as subconcepts of this role concept. This approach was motivated by the OntoClean approach (Guarino and Welty, 2002).

Based on inquiries among the VICODI users (historians, librarians) it seems that uncertainty and subjectivity appears mainly on different dimensions that can be modelled by traditional approaches like fuzzy and probabilistic logic (Baader et al., 2003). Although our users found also the features of those logics (like saying that someone is 'almost a King' or that he is 'a King with 0.8 probability') interesting, they felt that uncertainty and subjectivity are mainly present in the temporal and spatial dimensions. For example it is hard to define precisely the exact starting or ending time of the Middle Ages or the Russian Revolution. Imperfect spatial and temporal statements are hard to model with classical approaches, as those facts can only be expressed with many axioms together.

3 and other concepts like artifacts where it is meaningful to have roles

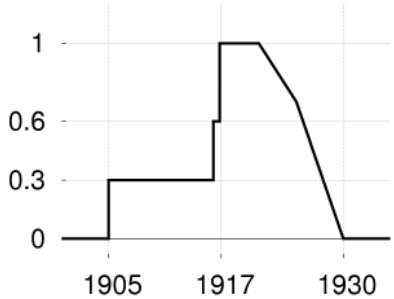


Figure 1. Existence time of 'Russian Revolution'

An alternative approach is using so called concrete domains[4]. Concrete domains are generally interesting from an ontological modelling point of view. Traditional logicbased formalisms are not very strong in handling strings, numbers etc. This aspect can be solved by introducing concrete domains (or datatypes) to the ontology language which allows us to 'plugin' different domainspecific reasoners. One usually does not want to logically infer that '1 + 1 = 2', it is enough, if this result is calculated by a suitable algorithm. This aspect is already realized by the research community, and concrete domains are part of the recent languages for the Semantic Web, such as OWL (PatelSchneider et al., 2004).

As the version of KAON which we used throughout the project did not support concrete domains, we had many problems especially with time modelling. *E.g.* retrieving all of the instances which are connected with a specific instance within a specific time frame – probably the most common operation in VICODI – was simply not possible without enumerating all of the connections and checking the condition 'manually' in application program code. This is clearly not the way ontologies should be used, and therefore – partially because of the problems we reported in VICODI – support for concrete domains is under implementation in KAON2[5], the successor of the KAON framework.

To deal with uncertainty and subjectivity, we designed a novel temporal model based on fuzzy logic, which is capable of representing the uncertainty and subjectivity aspects in a unified manner. The basic idea is to represent a time interval as a fuzzy subset of all possible temporal points, where the membership function of a specific point shows our belief that that very point belongs to the fuzzy interval. Fig. 1 shows a possible representation of the existence time of the Russian Revolution instance. A detailed description of this model is can be found in (Nagyp´al and Motik, 2003).

Although this model was not implemented during the VICODI project, an implementation using the new concrete domains feature of the KAON2 system is under development.

5 Methodological challenges

5.1 Solving the model mismatch

During the VICODI project the work of computer scientists (ontology engineers and specialists creating contextualization heuristics) and historians (domain experts) had to be coordinated. The biggest problem that we encountered during ontology development was how to synchronize the conceptualization of the two groups of experts. Computer scientists have significant knowledge about ontology formalisms, programming languages and algorithm design, but have only basic everyday knowledge about history.

Historians are naturally experts in history, but they felt formal ontologies unintuitive and cumbersome. The major technical cause for that was the fact, that RDFbased ontology formalisms (among them KAON and OWL (PatelSchneider et al., 2004)) support only binary relations, while time dependency in history renders practically all relations at least ternary (see Section 4). Although reification is a standard solution for such cases, it is not intuitive for the domain experts, which hinders the knowledge acquisition process significantly.

As an example consider the fact 'Napoleon was an emperor in France from 1804 to 1814'. The ontological representation of this fact is shown on Fig. 2. This representation is not directly intuitive for domain experts, and therefore they cannot encode the required facts into the ontology without permanent guidance by the ontology engineer. As a solution we applied the principle of intermediate models, which was already

4 In history we are not interested in relative temporal statements used by various temporal logics, but rather in temporal facts using concrete dates.

5 http://kaon2.semanticweb.org

successfully applied in other ontology development projects and methodologies (Lopez et al., 1999; Rector et al., 2001).

In our case we represented facts in Excel sheets, where the structure of the sheet was tailored to be optimal for a specific knowledge acquisition task. For example we had sheets for uploading instances together with their existence times, uploading instancerole relationships together with their validity times, uploading locationlocation relations with their validity times and so on. *E.g.* Table 1 shows a snippet from the Excel sheet defining person roles showing the mentioned fact about Napoleon.

Table 1. Representation of 'Napoleon is emperor' in the Excel file Instance

Instance	Role	Start year	End year	Location
Napoleon	Emporer	1804	1814	France

Those simple intermediate models turned out to be invaluable, and speeded up the knowledge acquisition process significantly. First, their structure were simple, and understandable for domain experts. Second, it turned out that the visual ontology editing paradigm supported by tools like Protégé[6] or KAON OIModeler[7] is absolutely not suitable for entering a mass amount of data. Features that legacy spreadsheet tools like Excel provide, such as easy search and replace and mass copypaste, are crucial for user acceptance. In addition the performance (in terms of response time) of established spreadsheet tools is much better than the performance of present ontology editors when dealing with big ontologies.

Of course, Excel does not support the necessary validation when entering new data what visual, ontology-aware tools provide. During the transformation process, however, when the Excel representation was uploaded into the KAON ontology, the various checks could be done based on the ontology structure, and improper facts could be refused. This also had the advantage that users could see the list of actual errors together, and correct them more effectively.

The huge success of our approach of allowing users to use the tools they were familiar with shows that it is a way which could be followed in many other ontology projects. This approach can probably solve the

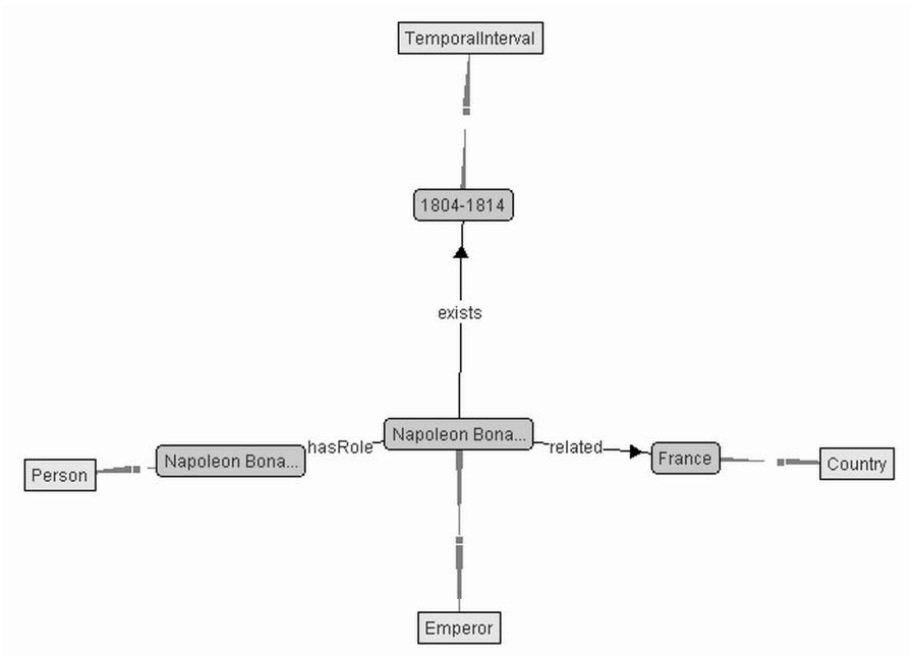


Figure 2. Representation of 'Napoleon is emperor' in the ontology

6 http://protege.stanford.edu/
7 http://kaon.semanticweb.org/

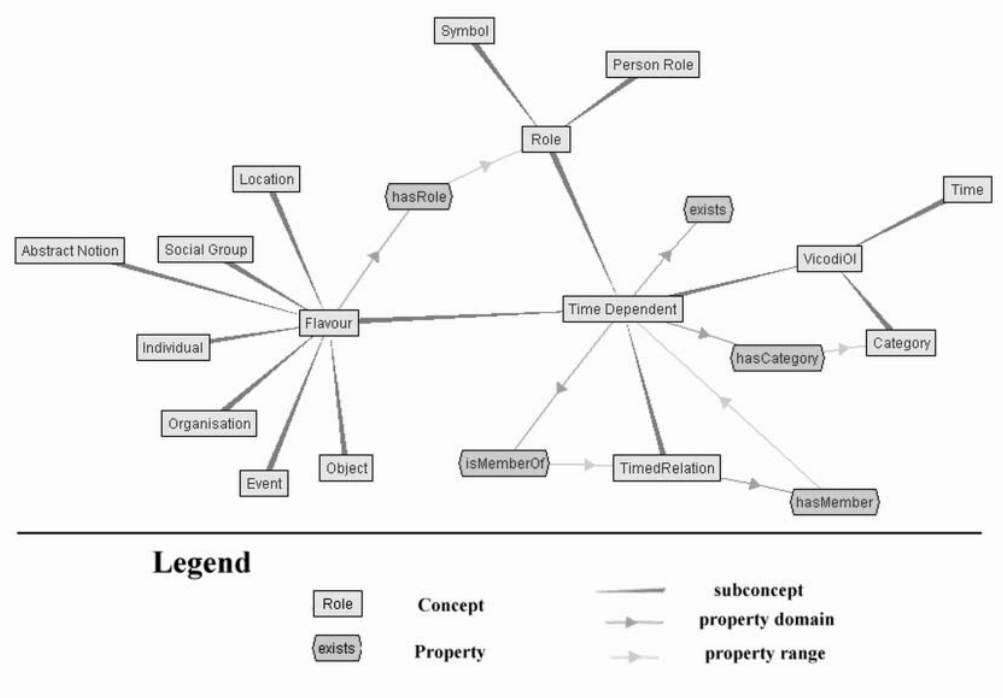


Figure 3. Excerpt from the highlevel structure of the VICODI ontology

usual dilemma: who should develop ontologies, domain experts or ontology engineers? Users encoding their domain knowledge in simple models using their favourite tools and ontology engineers refining those models into formal ontologies – this seems to be an economical solution.

5.2 Balance between simplicity and usability

In the early phases of the ontology development it turned out that either the concept or the property structure (or both) became unmanageable in a short time when our domain experts tried to model all subtle aspects of history. Even when we fixed the aspect we used to create the taxonomy of a flavor[8] (like individuals or events) it was very hard to decide where to stop with the taxonomy, because from a historical point of view even subtle details are of importance.

Therefore our domain experts generated very deep hierarchies describing the historical subdomain they were expert in. These deep hierarchies contained a huge number of concepts and were not intuitive for other experts. The same was true for the generated set of properties.

This caused many problems. First, no consensus could be reached among our domain experts, we were stuck in philosophical debates. Second, the complicated structure also made it hard for the experts to locate the proper place of new instances in the ontology. This slowed down the instance upload process significantly. Third, researchers designing algorithmic heuristics for context generation and context search did not have an overview over the ontology which made their work impossible.

Finally it turned out that it is not possible to design an ontology which encodes all of the subtle aspects of history but still remains understandable for all parties. As our main goal was to support the contextualization process (see Section 2), simplicity and extensibility (in terms of new instances) of the ontology was crucial. We therefore had to make a compromise and keep the ontology at a general level.

We achieved this by constraining the allowed levels of the created taxonomies – a technique used to deal with the complexity of the application domain *e.g.* in product classification standards like EClass[9]. This resulted in a structure which is although shallow and

8 Main concepts in the ontology were termed as 'flavors' in VICODI
9 http://www.eclass.de/

overly simplistic from a historian's point of view, is nonetheless intuitive, simple and allowed us to upload instances and relations (representing historical facts) into the ontology in huge numbers.

This approach turned out to be adequate for the purposes of VICODI, because the algorithms exploited only the highlevel concepts and properties during context generation and matching, and therefore they needed lots of links between ontology instances and lots of temporal facts about them, and a deep taxonomy was not beneficial for them. This will probably be the case for most of the applications employing machine algorithms to handle historyspecific information. Therefore the VICODI ontology can be a valuable asset in future historyrelated projects, as well. The present ontology has around 170 concepts, 15 properties and more than 15000 instances. Figure 3 shows an example of its high level structure.

Our experience shows that it is practically impossible to build a monolithic ontology for such a complex domain as history which reflects all of the subtle details of the domain. New techniques like ontology normalization (Rector, 2003) can perhaps help solve the problem by allowing the ontology to split into smaller, more manageable modules.

6 Conclusion

This paper presented our insights from the ontology engineer point of view which were gained during the development of an ontology of European history in the VICODI project. The most important lessons we learned were that pure ontological formalisms are not adequate to represent all of the intricate details of a complex application domain like history and that simple, intermediate models which are understandable for domain experts and can be manipulated by wellknown legacy tools, have a great potential to make ontology engineering more effective.

Although our experiences were specific to the domain of history, we made an attempt to generalize the lessons learned to provide some insights which – we believe – are useful for ontology development in social sciences in general.

References

Baader,F. , R. Küsters, and F. Wolter. *Extensions to Description Logics*, chapter 6, pages 219-261. Cambridge University Press, 2003.

Brickley, D. and R. Guha. Rdf vocabulary description language 1.0: Rdf schema. Recommendation, World Wide Web Consortium, 2004.

Deswarte, R. and J. Oosthoek. Clios ontology criteria: The theory and experience of building a history ontology. In *Proceedings of XVIth International Conference of the Association for History and Computing,* Amsterdam, the Netherlands, Sept. 14-17 2005a.

Deswarte R. and J. Oosthoek. Clio the difficult muse: Ontology challenges specific to the nature of history. In P*roceedings of XVIth International Conference of the Association for History and Computing,* Amsterdam, the Netherlands, Sept. 14–17 2005b.

Guarino, N., Formal ontology and information systems. In N. Guarino, editor, *Formal Ontology in Information Systems. Proceedings of FOIS'98*, pages 3-15, Trento, Italy, June 6-8 1998. IOS Press.

Guarino, N. and C. Welty. Evaluating ontological decisions with OntoClean. *Communications of the ACM*, 45(2):61–65, Feb. 2002. ISSN 00010782.

Lopez,M. F., A. GomezPerez, J. P. Sierra, and A. P. Sierra. Building a chemical ontology using Methontology and the ontology design environment. *IEEE Intelligent Systems*, 14 (5):37-45, Jan./Feb. 1999.

Motik, B. A., Maedche, and R. Volz. A conceptual modeling approach for semanticsdriven enterprise applications. In *Proc. 1st Int'l Conf. on Ontologies, Databases and Application of Semantics (ODBASE2002)*, Oct. 2002.

Nagypál. G., Creating an applicationlevel ontology for the complex domain of history: Mission impossible? In *Proceedings of Lernen Wissensentdeckung Adaptivit*.at (LWA 2004), FGWM 2004 Workshop, pages 287-294, Berlin, Germany, Oct. 4-6 2004.

Nagypál, G. and B. Motik. A fuzzy model for representing uncertain, subjective, and vague temporal knowledge in ontologies. In R. Meersman, Z. Tari, and D. C. Schmidt, editors, *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE,* volume 2888 / 2003 of Lecture Notes in Computer Science, pages 906-923. Springer Verlag, 2003.

Patel-Schneider,P. F., P. Hayes, and I. Horrocks. *Web Ontology Language (OWL) Syntax and Semantics*. W3C Recommendation, Feb.10 2004. URL http://www.w3.org/ TR/owlsemantics/.

Rector, A. L., Modularisation of domain ontologies implemented in description logics and related formalisms including owl. In *Proceedings of the international conference on Knowledge capture,* pages 121-128. ACM Press, 2003. ISBN 1581135831. doi: http://doi.acm.org/10.1145/945645.945664.

Rector,A. L., C. Wroe, J. Rogers, and A. Roberts. Untangling taxonomies and relationships: personal and practical problems in loosely coupled development of large ontologies. In *Proceedings of the international conference on Knowledge capture*, pages 139-146. ACM Press, 2001. ISBN 158113380 4. doi: http://doi.acm.org/10.1145/500737.500760.

Surányi,G. M., G. Nagypál and A. Schmidt. *Intelligent Retrieval of Digital Resources by Exploiting Their Semantic Context*, volume 3290. Springer, Oct. 2004. URL http:// springerlink.metapress.com/openurl.asp?genre=article&id=URP3VYAXB50NJ649.

Robust audio indexing for Dutch spoken word collections

Roeland Ordelman, Franciska de Jong, Marijn Huijbregts & David van Leeuwen***

Whereas the growth of storage capacity is in accordance with widely acknowledged predictions, the possibilities to index and access the archives created is lagging behind. This is especially the case in the oral history domain and much of the rich content in these collections runs the risk to remain inaccessible for lack of robust search technologies. This paper addresses the history and development of robust audio indexing technology for searching Dutch spoken word collections and compares Dutch audio indexing in the well-studied broadcast news domain with an oralhistory casestudy. It is concluded that despite significant advances in Dutch audio indexing technology and demonstrated applicability in several domains, further research is indispensable for successful automatic disclosure of spoken word collections.

1 Introduction

The number of digital spoken word collections is growing rapidly. Due to the ever declining costs of recording audio and video, and due to improved preservation technology huge data sets are created, both by professionals at various types of organisations and non-professionals at home and underway. Partly because of initiatives for retrospective digitisation, datagrowth is also a trend in historical archives. These archives deserve special attention because they represent cultural heritage: a type of content which is rich in terms of cultural value, but has a less obvious economical value. Spoken word archives belong to the domain of what is often called oral history: recordings of spoken interviews and testimonies on diverging topics such as retrospective narratives, eye witness reports, historical site descriptions, and modern variants such as 'Podcasts' and socalled amateur (audio/video) news[1].

Where the growth of storage capacity is in accordance with widely acknowledged predictions, the possibilities to index and access the archives created is lagging behind though [4]. Individuals and many organisations, often do not have the resources to apply even some basic form of archiving. Spoken word collections may become the stepchild of an archive – minimally managed, poorly preserved, and hardly accessible. The potentially rich content in these collections risk to remain inaccessible.

For 'MyLifeBits' chronicles collected by non-professionals under uncontrolled conditions [7] the resemblance with shoebox photo collections (i.e., little annotation and structure) may be acceptable. But for audio collections with a potential impact that is not limited to the individual who happened to do the recording, there is a serious need for disclosure technology. Tools for presenting and browsing such collec-

*Department of EEMCS, Human Media Interaction University of Twente, Enschede, The Netherlands
** TNO Defence, Security and Safety Soesterberg, the Netherlands

1 'Podcasts' are homebrew radio shows covering personal interest items and can be viewed as the audio variant of a 'blog' which is basically a journal that is made available on the web. Amateur news is news compiled by amateurs and broadcasted via the web.

tions and to search for fragments could support the information need of various different types of users, including archivists, information analysts, researchers, producers of new content, general public, etc.

The observation that audio mining technology can contribute to the disclosure of spoken word archives has been made many times [8], and several initiatives have been undertaken to develop this technology for audio collections in the cultural heritage domain. Worthwhile mentioning are projects such as ECHO (European CHronicles Online), that focused on the development of speech recognition for historical film archives for a number of European languages [1], and MALACH, applying ASR and NLP for the disclosure of holocaust testimonies [3]. But the high expenses required to process historical material in combination with the expected limited financial return on investment have prohibited real successes. A break through for the application of audio mining outside standard domains (typically: English news) is still pending.

This paper addresses the history and development of robust audio indexing technology for Dutch spoken word collections in various domains, including radio and television broadcasts, governmental proceedings, historical archives, lectures and meetings. The application of audio indexing technology for the oral history domain will receive special attention. Audio indexing involves topics such as audio partitioning, keyword spotting, speech recognition, speaker identification and information extraction. After introducing the technology in section II, this paper compares Dutch audio indexing in the wellstudied broadcast news domain (section III) with an oralhistory casestudy, consisting of a collection of spoken audio material from the Dutch novelist Willem Frederik Hermans (section IV). It is concluded in section V that despite significant advances in Dutch audio indexing technology and demonstrated applicability in several domains, further research is indispensable for successful automatic disclosure of spoken word collections.

2 Audio mining overview

Audio mining involves a number of research areas that have in common that they aim at the automatic extraction of information from audio documents that can directly or indirectly be used for searching. The extracted information can be regarded as document features; each feature adds to the overall representation of a document. Lowlevel and highlevel features are distinguished. Lowlevel features are for example acoustic features such as note duration and pitch in musical audio mining [9], or bandwidth and spectral features in speech mining. Instrument classifications, speech transcripts or even textual summaries are examples of highlevel features. The main focus in this paper is on the extraction of features that are relevant for generating higherlevel speech related features. Traditionally these include the localisation of the speech fragments (audio partitioning/segmentation), the speaker (identification and clustering) and the speech itself (speech recognition, information extraction). More recently the extraction of emotional features in the audio signal (*e.g.*, affect bursts such as laughter or words that express emotions), for example to detect so called 'hot spots' in collections, has been added to this list. Recent years have shown large improvements in the performance of automatic speech recognition (ASR) systems, and speech transcripts can now be generated at nearly the same quality as manual transcripts, at least for wellstudied domains such as the broadcast news domain [5]. As speech

recognition systems label recognised words with exact time information as a standard accessory, the timelabelled speech transcripts can directly be used to search within audio documents. Parts of a document that are relevant for a query can be accessed by linking to the timelabel of relevant words.

Before we give an overview of the most frequently applied speech recognition techniques in audio mining in section IIB, fundamental auxiliary techniques for a successful application of speech technology, audio segmentation and audio source labelling, are discussed in brief in section IIA. In section IIC, important issues for the presentation of search results for audio/video collections will be addressed.

A. Segmentation and audio source labelling

Although timelabelled speech transcripts can directly be used to search within audio documents, in practice longer audio documents are often prestructured: segmented according to a particular condition such as speakerturns, silence, or even topic, into homogeneous subdocuments that can be accessed individually. This is convenient, as scrolling through a large unstructured audio or video document to identify interesting parts can be cumbersome. Audio segmentation can be advantageous from a speech recognition point of view as well, as it allows for segment based adapta-

tion of the recognition models as will be discussed below. A frequently applied adaptation scheme is based on speaker identity.

Using a fixed overlapping time window, or fixed number of words to segment an audio stream is a simple but in cases very effective segmentation approach that does not rely on special segmentation tools. When the window and overlap ranges are chosen well, it can provide a document structure that can already usefully be deployed for certain retrieval tasks, such as wordspotting. But a segmentation based on audio features is much more informative and helpful both from a retrieval and speech recognition point of view. With a segmentation according to speaker for example, retrieval results can be structured and presented according to speaker identity. In addition speaker dependent modelling schemes can be applied in order to improve speech recognition performance. Useful segmentation cues are in general provided by techniques that aim at the labelling of the source of audio data (*e.g.*, acoustic environment, bandwidth, speaker, gender), often referred to as 'diarisation' or 'nonlexical information generation' [18].

From a retrieval point of view, topicbased audio segmentation would be an obvious choice. Such a segmentation would resemble the topic structure in textual documents. In addition, topic segmentation allows for the selection of topic specific language models and rescoring the speech recognition results with these topicspecific models. Topic segmentation can for example be based on word frequency and cooccurrence information (see *e.g.*, [15]).

B. Speech recognition

Information retrieval research that uses the spoken audio parts of documents for retrieval is commonly referred to as spoken document retrieval (SDR) or alternatively, speechbased retrieval. Recent years have shown that automatic speech recognition can successfully be deployed for equipping spoken word collections with search functionality. This is especially the case in domains such as the broadcast news domain which is very general and makes data collection for system training relatively easy. For the broadcast news domain speech transcripts therefore approximate the quality of manual transcripts for several languages and spoken document retrieval in the AmericanEnglish broadcast news (BN) domain was even declared 'solved' with the NIST sponsored TREC SDR track in 2000 [5]. In other domains than broadcast news, a similar recognition performance is usually harder to obtain due a lack of domain specific training data, in addition to a large variability in audio quality, speech characteristics and topics that are addressed. This applies to the oralhistory domain in particular.

The most obvious approach in spoken document retrieval is the wordbyword translation of the encountered speech using a large vocabulary continuous speech recognition (LVCSR) system. Having generated a textual representation (full text transcription) of an audio or video document, the document can be searched as if it were a text document. As mentioned above, the timelabels provided by the speech recognition system and the segmentation of a large document into subdocuments, provide additional means for structuring the document.

For some applications, applying a wordspotting approach can be beneficial as an alternative or auxiliary tool. This is especially the case when the mismatch between speech recognition vocabulary and domain vocabulary is hard to model and tends to produce many outofvocabulary words. That is, when in a certain domain very specific words are used frequently (*e.g.*, jargon, names), and there are no resources to rely on for the prediction of these words (*e.g.*, related text documents). As a consequence, the speech recognition vocabulary will not 'know' the specific words and will not be able to recognise them. This implies that when these words are used as search terms the retrieval system will not find the documents that contain this term. In addition, as the 'unknown' word is replaced by the speech recognition system for a substitute word the system knows (and is acoustically similar to the unknown word), the wrong documents will be produced by the retrieval system when a query contains exactly these substitute words.

A keyword spotter searches the audio material for single keywords. This can be done 'online', immediately after a user has posed a query, but this requires that the archive is not too large as it takes some time to do the online processing. In an 'off-line' approach, the word spotter searches for the appearance of a relatively small set of keywords in the documents in advance. The words that are found are then used to label the document. An acoustic model is used to recognise phones and a, usually small, vocabulary of keywords with phonetic transcriptions provides the link to the keywords.

In specific cases, speechbased retrieval can do without speech recognition technology. In the ORL Medu-

sa multimedia retrieval system [2] for example, teletext subtitles were exploited for purely speechbased retrieval. Subtitles contain a nearly complete transcription of the words spoken in the material and provide an excellent information source for indexing when aligned (labelled with timecodes) to the actual speech in the collection. A comparable strategy was used for disclosing meetings of the Dutch Government. Here, instead of subtitles, the official stenographic minutes ('Handelingen van de Tweede Kamer') that closely follow the discourse of the meeting, have been aligned to the speech. When available, deploying such external information sources can be a practical solution for certain task domains. If necessary, speech recognition technology may still be called upon to timealign the external transcripts with the speech in the documents.

C. Presentation of search results

The presentation of search results for a given collection in spoken document retrieval deserves special attention. People are generally very good at browsing globally through text documents, and often recognise at a single glance whether a document is relevant for their specific information need or not. For audio/visual documents, browsing globally is not an option; one usually has to play the entire document to be able to assess its relevance. Providing fastplay functionality can sometimes be helpful to speed up the process.

By structuring an audio/visual document via the deployment of segmentation techniques for the creation of subdocuments, presentation functionality can be improved considerably. A speaker based segmentation for example, makes it possible at least to skip certain speakers. In television broadcast news shows, anchor people pause longer between topics then within topics. Here, a simple pause based segmentation thus provides a helpful topic structure for browsing.

Providing the automatically generated speech transcripts with the (sub)documents for rapid relevance assessment can be an option. However, the raw speech recognition transcripts contains errors, and, as punctuation is absent, can be hard to read. In addition, distracted by the speech recognition errors in the transcripts, a user may be prejudiced concerning the relevance of results. Research aiming at the generation of readable transcriptions incorporating capitalisation, punctuation, and speaker markers is currently investigated among others in the DARPAEARS Rich Transcription project[2]. The speech transcripts can still be used for contentbased topic segmentation (*e.g.*, in less structured domains) and automatic summary generation, provided that the speech recognition transcripts are reasonably accurate.

A relatively novel approach is the linking of collection fragments to external sources of information [10]. For example, fragments of digital meeting recordings can be linked to meetings agendas or policy documents pertaining to specific topics of a meeting. Such documents can be considered as part of the same archive. However, for a meeting archive links to external sources may also be useful, such as newspaper articles, an archive of broadcast news items and documentaries, and a 'Who-is-Who' of politicians. Given the potential size of such external sources, and the variety of perspectives that can be taken on them, automatically generated links could be very beneficial. As the external information sources can have different media formats (text, audio, video) this type of disclosure is referred to as crossmedia disclosure. (For work on crossmedia disclosure for meeting recordings in the IST-project AMI, cf. [14].)

Another strategy to enhance the presentation of search results for audio/video collections is making use of visualisations. For example, the entire audio document can be represented by a bar and fragments in the audio document that match a query are then represented within the bar as red markers. These markers may vary in length or colourdepth to represent the degree of relevance of the particular fragments. By pointing at the markers with the mouse, the audio fragment can be listened to.

3 Dutch spoken document retrieval

Spoken document retrieval research for Dutch first targeted at the development of a Dutch LVCSR system for the broadcast news domain. Broadcast news (BN) data has been extensively used as a benchmark domain for both international speech recognition research and SDR research. The domain is wellstructured and a lot of resources exist that can be explored for the development of a broadcast news recognition system (*e.g.*, newspaper data for language modelling, autocues for acoustic modelling). In the next section, the Dutch broadcast news speech recognition system is

2 The Effective Affordable Reusable SpeechToText (EARS) program: http://www.darpa.mil/ipto/programs/ ears

described and the results of a Dutch broadcast news retrieval experiment are discussed.

A. Broadcast news system

In [11] research on Dutch spoken document retrieval in the broadcast news domain is described in detail. In [11], the ABBOT speech recognition system [16] was used for the generation of speech transcripts. Recently the Dutch models have been ported to a new speech recognition system (referred to as UTBN2005 system), that is based on a recogniser developed at the University of Colorado (CSLR) which has been made available for research purposes. Its acoustic models are decisiontree stateclustered Hidden Markov Models [12] and the broadcast news specific models are trained using twentytwo hours of broadcast news recordings from the Spoken Dutch Corpus ([17]). It uses a large vocabulary of 65 K (65 thousand) words and a statistical trigram language model derived from a Dutch newspaper corpus of some 400 M (400 million) words (referred to as TWente News Corpus). On a broadcast news test set consisting of 4 hours of broadcast news material from the Spoken Dutch Corpus, the UTBN2005 obtained a word error rate (WER) of 30 %. Although at TREC8, BN systems for American-English produced error rates below 20 %, this figure can be considered an adequate baseline given that it concerns a relatively simple (unadapted, singlepass) system trained on a medium size acoustic training data set and with a standard 65 K newspaper vocabulary and language model.

To investigate the applicability of the developed Dutch speech recognition system in a retrieval task, a knownitem retrieval task that simulates a user seeking one particular document [19], was performed. We used a set of 18 television news broadcasts (*NOS Acht uur journaal*) that were segmented manually: 180 stories with a mean length of 257 words. Introductions and weather reports were excluded. Story topics were generated by students who were instructed to create topic 'titles' that in a few words (with a maximum of ten words) give a reasonable impression of the contents of the story. These titles were further interpreted as queries aiming at the retrieval of the respective stories. The titles were used as queries for retrieval given the following evaluation modes: using document representations that are based upon perfect, human-transcribed reference, using document representations based upon a speech recognition system producing a relatively large number of errors (word error rate of 50 %) and representations based upon a relatively well performing speech recognition system (word error rate of 30 %).

Table I shows the results of the knownitem retrieval task. Using the reference transcript as document representation gave the best retrieval performance in terms of found documents. Using high quality speech recognition produced comparable results: only one document less was found (10 instead of 9) and the mean rank when found was even slightly better compared to the one obtained in the reference condition. As could be expected, deploying a low quality speech recognition system significantly worsens retrieval performance. Almost a quarter of the documents could not be found, on average, the target stories were retrieved almost one rank lower, and the mean reciprocal rank decreased almost 35 % relative compared to the high quality speech recognition condition.

Table 1 Results of the knownitem retrieval task: mean rank when found (MRWF) and number of documents not found.

document representation	MRWF	not found
Reference	2.0778	9 (5 %)
ASR with low WER	1.9278	10 (5.6 %)
ASR with high WER	2.8556	43 (23.9 %)

This retrieval experiment showed that for spoken retrieval in the broadcast news domain, a speech recognition accuracy of some 70 % produces similar retrieval results as manually generated speech transcripts. When recognition accuracy decreases to 50 % however, retrieval performance drops considerably. This is in accordance with figures seen in the TREC SDR evaluations and in a Dutch spoken document retrieval simulation study as described in [20].

4 Willem Frederik Hermans case study

First explorations outside the broadcast news domain made clear that for other domains there is still a lot to accomplish, especially for the accuracy level of ASR. Experiments with the ECHO historical archive collection learned that search technology based on ASR might easily collapse due to shockingly high word error rates caused by the typical characteristics of historical material, for example: a wide variety in audio quality, background noise, overlapping speech, spontaneous

speech, topics that are unknown beforehand, old fashioned speech, dialect speech.

In the case study that will be described below, the target is on a slightly more heterogeneous oral history collection with (almost) only one speaker: lectures and interviews of the wellknown Dutch novelist Willem Frederik Hermans (1921–1994). The collection can be searched via the Willem Frederik Hermans webportal[3]

Although the performance of a BN system in the oral history domain was expected to be poor, we used the tools and resources collected and developed for a broadcast news (BN) system as a starting point. As similar systems are available in many labs, the conversion of the BN system and tuning to a collection from the oral history domain might be a case of a more general interest for research groups that want to pursue applications for their ASR tools for similar purposes.

Below we will first describe the Willem frederik Hermans audio collection (IVA) and the data that was used to train the speech recognition system (IVB). Next, we will compare the speech recognition performance on the collection of the broadcast news system and a system that is tuned to the domain (IVC).

A. The collection

The collection of audio recordings to be disclosed consists of some 10 to 15 hours of lectures and interviews featuring Willem Frederik Hermans (WFH). More data will become available at a later stage. Although WFH is not the only speaker present in the material, his voice dominates the larger part of the collection. The lectures were recorded at different locations with different acoustics. The lectures which were studied in more detail have applause, laughter, coughing and questions from the audience that – even for a human listener – sometimes are hard to recognise. Parts of the interviews are quite informal and recorded in a home environment on celluloid tape.

B. Training data

One of the lectures and a television documentary with a number of interviews were manually annotated at word level (130 minutes of speech), with WFH speaking approximately 85 % of the time. A training set (78 minutes) was used for training the acoustic models. A test set was used to evaluate both the acoustic models and the language models during development. An evaluation set was used for the final evaluation of the system.

In the annotated speech material, WFH is speaking 110 out of the 130 minutes. It is therefore reasonable to expect that the recognition rate will improve when a speaker (WFH) dependent acoustic model is used instead of the broadcast news acoustic model. Two new acoustic models were trained. The first model was trained solely on the part of the training set in which WFH is speaking. The second model was created by adapting the broadcast news acoustic model (UTBN2005) to the training data using an acoustic adaptation algorithm referred to as Structured Maximum a Posterior Linear Regression (SMAPLR) adaptation [12].

For training BN language models we used the Dutch newspaper corpus. Two other text collections were available for domain adaptation. A number of written interviews with WFH and one of his short novels made up the first collection (further referred to as WFHtext) containing one and a half million words. Wordlevel transcripts of general conversational speech from the Spoken Dutch Corpus [17] formed the other text collection. This collection consists of 1.65 M words. Both text collections were used to apply language model adaptation schemes in order to create a language model that better fits the WFH domain.

Two domain specific trigram language models were trained. These models both use a 30 K vocabulary containing domain specific words. The most occurring words from the WFHtext described above were complemented with the most occurring words from the newspaper corpus.

From each of the two domain specific text collections a language model was created. A third model was created using the newspaper corpus and the 30K vocabulary. From these three models, a mixture language model that combines the statistics of each separate language model into a single language model using a weight factor was created. Mixture weights were computed using the transcripts of the acoustic training set.

C. Experimental results

The word error rates of the broadcast news acoustic model and the adapted acoustic models on the WFH-collection are shown in Table II. In order to make a fair

3 http://www.willemfrederikhermans.nl/

comparison with the broadcast news system, the 65 K broadcast news language model was used during these recognition runs. Table II shows three word error rates for each model. The first WER is of the part of the audio in which WFH is speaking, the second one is based on speech from other people and the third is the overall word error rate.

Both adapted models perform better than the broadcast news model. Although the broadcast news model performs best on the small subset with various speakers (15 % of the total amount of speech), the adapted models show improved WERs on the part of the data in which WFH is speaking. The SMAPLR adapted model (66.9 % WER) outperforms the speaker dependent model (76.6 % WER). Apparently, the 78 minutes of speech used for training the speaker dependent model does not contain enough data for building a robust acoustic model.

Table 2. WERs of three acoustic models: the 2005 broadcast model, the WFH model and the SMAPLR adapted model. The second column shows the WER of the part in which WFH is speaking, the third the WER on the other speech parts of the evaluation set. The last column shows the total WER.

	%WER AM	%WER WFH	%WER other total
UT-BN2005	81.6	67.2	80.4
WFH	76.0	83.2	76.6
BN-SMAPLR	66.7	77.1	67.5

The speaker adaptation we employed here, is a socalled 'supervised adaptation.' The segmentation into speakers (WFH and nonWFH) has been performed manually, and the acoustic models have been adapted to the speaker WFH using transcribed text. Both the segmentation and the speaker adaptation can in principle be performed automatically, or 'unsupervised.' For speaker segmentation an acoustical segmentation/clustering algorithm can partition the audio stream in segments belonging to the same speaker [6]. Using a first speech recognition pass, an automatic transcript can be generated. For each cluster of audiosegments spoken by the same speaker, this automatic transcript can be used to adapt the speakerindependent acoustic models to clusterdependent models. These models can then be employed for a second speech recognition pass for the speech segments in that cluster, in order to obtain more accurate transcripts. Our supervised SMAPLR experiments give an impression of the maximum achievable performance increase, reducing the WER from 81.6 % to 66.7 %, for speaker WFH.

In Table 3 the word error rates are shown of a system that uses both the adapted acoustic model and the domain specific 30K mixture language model. The combination of the 30K mixture language model and the SMAPLR adapted acoustic model results in the best system performance: 66.9 % WER.

Table 3. The word error rates (WER) of the two adapted acoustic models combined with the 30K mixture language model. The first row contains the AM trained on WFH solely. The second row contains the SMAPLR adapted acoustic model. For comparison, the results from the baseline BN language models are given as well.

AM	%WER WFH	%WER other	%WER mixLM	%WER BN-LM
WFH	73.8	83.6	74.6	76.6
ADP	66.4	72.5	66.9	67.5

By creating a mixture LM and a speaker dependent AM the word error rate was reduced with 13.5% (16.8% relative). To determine possible further improvements, a brief error analysis was conducted. To investigate to what extent different audio conditions influence the word error rate, speech segments were classified into five classes: clean speech (*F*0), speech with audible reverberation (*F*1), speech containing background music (*F*2), speech with background noise (*F*3) and overlapping speech (speech interrupted by other speakers, *F*4).

Table 4 shows the word error rates in each of the conditions. Two third of the segments in the WFH evaluation set are classified as 'clean.' One third contains reverberation, music, noise or overlapping speech. Although all speech in the music class is clearly understandable for human listeners music increases WER substantially, in the WFH task by more than 10%, which is comparable with the statistics reported in [13].

Table 4. The word error rates of the five manually classified parts of the WFH evaluation set.

Class	%WER WFH	%WER other	%WER total
F_0	63.9	61.8	63.8
F_1	75.4	100	76.5
F_2	76.1	86.1	78.6
F_3	82.4	83.3	82.4
F_4	100	100	100

5 Discussion and conclusion

The Willem Frederik Hermans case study revealed that simply deploying the broadcast news system for transcribing oral history data resulted in high error rates of around 80% WER. In order to improve on the BN system, several adaptation schemes were applied on the acoustic level and on the language model level that indicated that a maximum achievable performance increase, reducing the WER from 81.6 % to 66.7 % for speaker WFH, is achievable. The overall word error rate was 66.9%, a 13.5% absolute improvement on the baseline BN system. Although near perfect transcripts are not required for a successful application of spoken document retrieval, error rates in this range are clearly below threshold.

It has already been noted that the large variability in audio quality, speech characteristics and topics are typical for the oral history domain and make the successful application of speech recognition technology difficult. A descriptive study on the characteristics of a certain collection is an important minimal prerequisite for identifying useful developments strategies.

What makes a successful application even more difficult is the fact that for the oral history domains we have seen until now, related audio and text sources that could be used for adapting the speech recognition components to these domain characteristics are usually only minimally available. This can be due to the fact that oralhistory archives have limited resources so that links to useful metadata are simply missing, or to the historical nature of the collections. For example, in the ECHO collection we could only use contemporary newspaper texts to model the ancient, outdated speech of the Dutch Queen Wilhelmina (18801962), as there were no example text data digitally available that could be used to model this type of speech. An attempt to apply OCR techniques on related historical text data failed because of the low quality of the paper copies. Next to text (language model) related problems, ancient or dialectic speech that does not or only minimally occur in contemporary speech training databases, impose additional constraints to the effort to obtain an adequate speech recognition performance. A first step towards the successful application of the automatic disclosure of oralhistory collections, should therefore be to collect (from a speech recognition developer point of view) or make available (from a content provider point of view) as much related data sources as possible for finetuning the system. A strategy to deal with the lack of acoustic training data is deploying (partly) unsupervised training strategies.

Other topics that need to be addressed are related to the retrieval functionality proper. Dependent on the users that are expected to search the collections, this functionality may need to be adapted. For example, users may use highly selective, domain specificwords in their queries that are infrequent in the material. Because of their low frequency, such words have a low chance of being selected for the speech recognition vocabulary and thus become so called 'query outofvocabulary'. Applying a word spotting approach to search for such words would then be an option. Another example concerns the presentation of the retrieval results. Presenting a user with short excerpts from the collection that contain query words may not be very informative. Instead, providing coherent fragments that are structured according to speaker or topic may be preferred.

It can be concluded that applying audio mining techniques for the disclosure of oralhistory collections is a promising approach. Proofofconcept has already been provided in other domains. However, due to the typical characteristics of the oral history domain, substantially more effort must be directed towards obtaining speech transcripts that can be used adequately for indexing. Research aiming at the optimisation of presentation strategies, of interest for spoken word collections in general, could also boost the usability of audio mining considerably.

Acknowledgement

This paper is based on research funded in part by the Dutch projects MultimediaN and Waterland. We like to thank the Willem Frederik Hermans Institute for providing text and audio material.

References

[1] ECHO Project Homepage. http://pcerato2. iei. pi.cnr.it/echo/.

[2] M. G. Brown, J.T. Foote, G.J.F. Jones, K. Sparck Jones, and S. J. Young. Automatic Contentbased Retrieval of Broadcast News. In *Proceedings of the third ACM international conference on Multimedia*, pages 35–43, San Francisco, November 1995. ACM Press.

[3] W. Byrne, D. Doermann, and M. Franz. Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives. *IEEE Transactions on Speech and Audio Processing, Special Issue on Spontaneous Speech Processing*, July 2004.

[4] K. W. Church. Speech and Language Processing: Where Have We Been and Where Are We Going? In *Eurospeech_2003*, Gen`
eve, Switzerland, September 2003.

[5] J.S. Garofolo, C.G.P. Auzanne, and E.M Voorhees. The TREC SDR Track: A Success Story. In *Eighth Text Retrieval Conference*, pages 107–129, Washington, 2000.

[6] JeanLuc Gauvain, Lori Lamel, and Gilles Adda. The LIMSI broadcast news transcription system. *SPeech Communication*, pages 89–108, 2002.

[7] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C.Wong. Mylifebits: fulfilling the memex vision. In *ACM Multimedia*, pages 235–238, 2002.

[8] Jerry Goldman et al. Report of the EU/NSF working group on Spoken Word Audio Archives. http:// www. ercim.org/publication/wsproceedings/ DelosNSF/Spoken word.pdf, 2003.

[9] A. Goodrum and E. Rasmussen. Sound and speech in information retrieval: an introduction. Bulletin of the American Society for Information Science, 2000. (URL: http://www.asis.org/Bulletin/ June00/godrumrasmussen.html).

[10] Jeroen Morang, Roeland Ordelman, Franciska de Jong, and Arjan van Hessen. InfoLink: analysis of Dutch broadcast news and crossmedia browsing. In *Proceedings of ICME 2005 (to appear)*, Amsterdam, September 2005.

[11] Roeland Ordelman. *Dutch Speech Recognition in Multimedia Information Retrieval*. PhD thesis, University of Twente, The Netherlands, October 2003.

[12] B. Pellom and K. Hacioglu. Recent Improvements in the CU Sonic ASR system for Noisy Speech: The SPINE Task. In *Proc. ICASSP*, 2003.

[13] B. Raj, V. N. Parikh, and R. M. Stern. The Effects Of Background Music On Speech Recognition Accuracy. In *Proc. of the ICASSP, Munich, Germany*, 1997.

[14] S. Renals. Ami: Augmented multiparty interaction. In *Proc. NIST Meeting Transcription Workshop*, Montreal, 2004. AMI 10.

[15] P. Rennert. Streamsage unsupervised asrbased topic segmentation. In *TRECVID 2003 Text REtrieval Conference Video Track*, Gaithersburg, Maryland, November 2003.

[16] Tony Robinson, Mike Hochberg, and Steve Renals. *The use of recurrent networks in continuous speech recognition*, chapter 7, pages 233–258. Kluwer Academic Publishers, 1996.

[17] I. Schuurman, M. Schouppe, H. Hoekstra, and T. van der Wouden. CGN, an Annotated Corpus of Spoken Dutch. In *In Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC03)*, 2003.

[18] S.E. Tranter, K. Yu, D.A. Reynolds, G. Evermann, D.Y. Kim, and P.C.Woodland. An investigation into the interactions between speaker diarisation systems and automatic speech transcription. Technical report, Cambridge University Engineering Department, October 2003.

[19] E. Voorhees, J. Garofolo, and K. Spärck Jones. The TREC6 Spoken Document Retrieval Track. In *Proceedings DARPA Speech Recognition Workshop*, 1997.

[20] E. Zuurbier. Onderzoek naar de haalbaarheid van spoken document retrieval. Master's thesis, University of Twente, 2004.

From the roman eagle to E.A.G.L.E.: harvesting the web for ancient epigraphy

Simonetta Pasqualis Dell Antonio

The international epigraphic database confederation EAGLE (Electronic Archives of Greek and Latin Epigraphy) and its portal want to be a news tool of investigation for researchers in the field of epigraphy and ancient history, with a long term aim of making all Latin and Greek inscriptions from antiquity available through internet in a standardized system of criteria. The project started in Rome in 1999 under the aegis of the Commission 'Epigraphie et Informatique' de l'Associaton Internationale d'Epigraphie Grecque et Latine (AIEGL), and in Trieste in 2003 was it decided to form a confederation of the already existing databases of Rome (EDR), Heidelberg (EDH), and Bari (EDB). The portal nowadays provides access to more than 15000 inscriptions from EDR (concerning the city of Rome and the Italian Peninsula), 36000 from EDH (concerning the Roman Provinces), 23000 from EDB (concerning the Christian epigraphies). This enormous epigraphic patrimony is now easily searchable through EAGLE so that scholars and institutions all over the world can enjoy a lively and fast exchange of information.

In the origin there were three databases.....

The oldest database dates back to 1986 when the Epigraphischen Datenbank Heidelberg was funded by the Gottfried-Wilhem-Leibneiz-Preises as a long term project for the registration and recording of Ancient Latin Inscriptions. At the same time it was decided to establish the Bibliography and the Photographic database as well.

In 1993 the EDH became a Research Project of the Heidelberg Akademie der Wissenschaften, and in 1997 it went online comprising all the inscriptions which had been entered up to that date (almost 30000). It gets regular updating every six months, and combined searches are possible with the help of a search engine.In 1997 during the 11th International Congress of Latin and Greek Epigraphy held in Rome EDH was presented. In 1999 the photographic material which accompanies those Imperial Inscriptions of Rome entered in EDH, was presented on Internet and in 2000 the photos from the Archive of the Spanish External Branch of the Corpus Inscriptionum Latinarum at the University of Alcalá were entered to EDH together with the accompanying inscription texts (www2.uah.es/imagines_cilii/). In 2001 inscriptions were entered from Algeria and Libya, and in 2003 all three databases belonging to the EDH were transferred to DB2: from this point onwards the entry of data became possible both on local and on world wide basis.

In 2003 a collaboration started with the members of the Institut für Alte Geschichte, Altertumskunde und Epigraphik of the University of Vienna (inscriptions from Noricum and from the Austrian part of Pannonia Superior). In 2004 the University of Graz joined (inscriptions form Carinthia), and in the same year the Epigraphic Bibliography (EBH) went online. In the near future there will be contributions by members of the Institute of Archaeology of the University of Lubljana.

Since November 2004 the Epigraphische Bibliographie (EBH) can be accessed on the web: it contains more than 10.000 records concerning monographs and articles in journals, anthologies, commemorative

volumes and Proceedings, and almost all this literature is present at the research center at the very least in the form of photocopy.

Another important tool of research is the Photographic Database, which provides over 11.000 photos from Rome and Hispanic Provinces, and at present almost as many photographs from all parts of Imperium Romanun. Alongside archiving of conventional photographic material such as slides, negatives and prints, new material is increasingly archived in digital form. Whenever there exists a corresponding record to the text of an inscription, the images are presented as clickable link with the text in a search result, or an indication is given that photographs of the inscriptions exist in the photographic library. The database EDH is available at the address http://www.uni-heidelberg.deinstitute/sonst/adw/edh (1)

The Epigraphic Database Bari (EDB) was conceived for the first time in 1988 with the aim to find a tool able to search through the enormous amount of Christian inscriptions of the city of Rome (Inscriptiones cristianae urbis Romae – ICVR) ranging from the 3rd to the 8th century AD. All inscriptions are collected of course in the many volumes of the ICVR nova series, which offer about 45000 inscriptions. The need to create an informatic database was evident and it was decided to present a project to the Council of the Pontificio Istituto di Archeologia Cristiana: on the basis of a MS-DOS programme for textual processing data entry was started, and the first results were presented on the occasion of the 7th National Congress of Christian Archaeology, at Cassino in 1993.

Also Greek inscriptions were put in the database, and in 1999 it was decided to join the programme created by the Commission 'Epigraphie et Informatique' of AIEGL and presented during the Meeting of Rome (28-29 May 1999). On that occasion it was decided to adopt a new font, called Epigraph, an adaptation of Code2000 made by the researcher Anita Rocco to put data in, in order to adopt a uniform format. (At that time font Code2000 was free for use and adaptation, but now this is no longer the case so that also Bari had to resort to font Cardo, the one used by EDR). Since 2003, after the International Workshop presenting EAGLE held in Trieste and Aquileia in November 2003, EDB is a federated database within the EAGLE project. Future developments will envisage:

– the input of Christian inscriptions not already covered by ICVR, in collaboration with the Pontificia Commissione di Archeologia Sacra, Fabbrica di San Pietro, Vatican Musea, and all other institutes responsible for conservation of these documents.
– the creation of a parallel database for digital images of the inscriptions, in agreement with the above mentioned institutes.
– the creation of a database for epigraphic material from the high medieval and medieval periods, and at the same time the creation of Corpus Inscriptiones Medi Aevi Italiae (IMAI) sponsored by the Centro Italiano Studi sull'Alto Medioevo (CISAM) of Spoleto.

This database is available at:www.edb.uniba.it (2)

The Epigraphic Database Roma (EDR) was started as an experimental database accomplishing the decisions taken in Rome in 1999 by the Commissione 'Epigraphie et Informatique' of AIEGL. From 1999 to 2003 the project was developed and carried out by a working group whose main members are: the course of Epigrafia Latina Facolta' di Scienze Umanistiche dell'Università di Roma La Sapienza, Centro Interdipartimentale di Servizi per l' Automazione delle Discipline Umanistiche La Sapienza (CISADU), Dipartimento di Studi Classici e Cristiani dell'Università di Bari, Epigraphische Datenbank Heidelberg (EDH), Laboratorio di Epigrafia Latina dell'Università di Trieste.

EDR presents now a tree structure whose core is in Rome, and within the EAGLE project EDR is responsible for the inscriptions coming from Rome, the Italian Peninsula, Sicily and Sardinia. The database enables to visualize its epigraphies through the Font Cardo, created by Dr. D.J. Perry of the Rye High School of New York. At the moment a new programme of cataloguing is in progress, and it will concern Rome and the following Regiones: I (Latium et Campania), II (Apulia et Calabria), IV (Sabina et Samnium), V (Picenum), VI (Umbria), IX (Liguria), X (Venetia et Histria), XI (Transpadana). Furthermore, it has been envisaged to integrate in EDR, with the permission of the author, the material once catalogued by prof. John Jory and collected in a CD-ROM called 'Epigraph. A database of Roman Inscriptions'. In the future there will be a Photographic database as that of EDH, but Rome is now waiting for the Ministry authorization to put their photos on the web. The Roman database is visible at www.edr-edr.it (3)

The first phase: Tituli Antiqui Collecti

On the occasion of the 11th International Congress of Greek and Latin Epigraphy, held in Rome in 1997, the Commission 'Epigraphie et Informatique' of AIEGL was renewed and for the term 1997-2002 the appointed members were: Silvio Panciera (President), Geza Alfoldy, Alain Bresson, Kevin Clinton, Charles V.Crowter, Manfred Hainzmann.

This Commission was given the following tasks:

a) to verify the possibility to store in a computerized way the texts of all Latin and Greek inscriptions, organizing the coordination of already existing data, and to promote short and long term cooperation of other epigraphic institutes
b) to agree upon a common format for transforming or putting in data
c) to establish criteria to be followed by all participants so that memorized data are consistent
d) to assess all existing programs having characteristics useful to accomplish step a)
e) to contact all scholars responsible for similar projects in order to start cooperation based on steps a-b-c) and to organize a statutory round table (4).

The aim of step d) was achieved by producing a booklet 'Greek and Latin Epigraphic Corpora: state of art and updating' edited by Silvio Panciera, 1999, which described 21 existing computer experiments. Step e) was accomplished by organizing the round table held in Rome 28th and 29th May 1999 promoted by Università di Roma La Sapienza and Ecole Francaise de Rome.

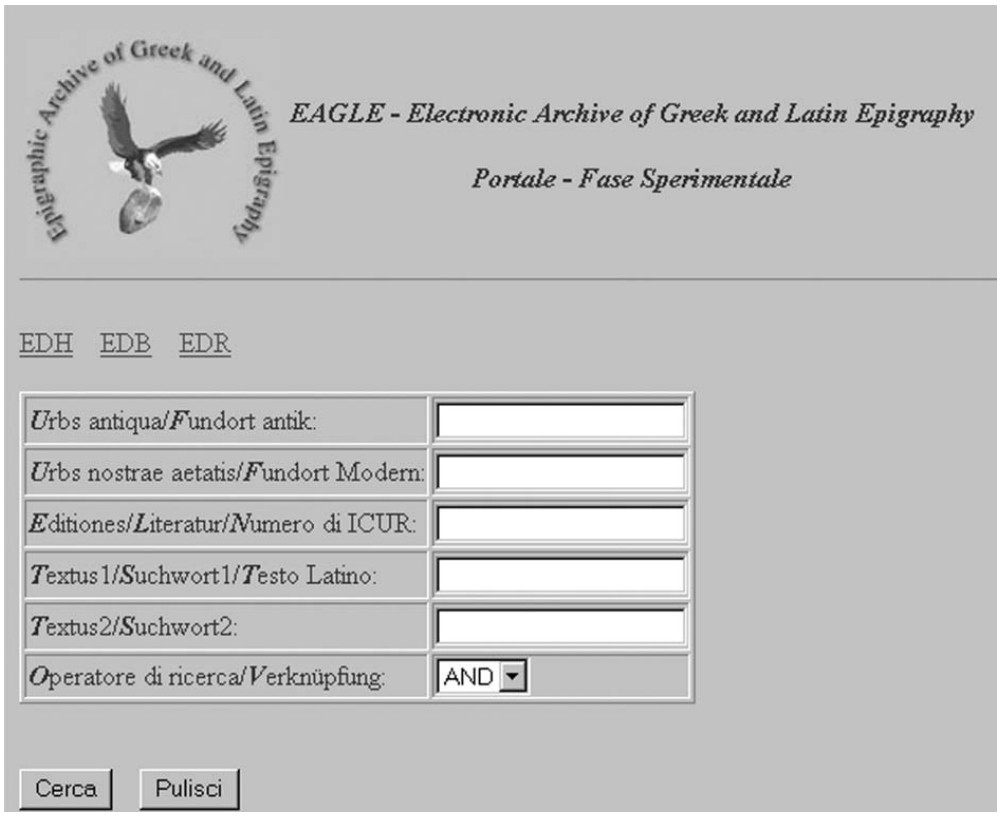
The Round Table gave rise to the following decisions:

1. the creation of a relational database where all Greek and Latin ancient inscriptions are registered according to the best edition
2. the name of the database will temporarily be TITULI ANTIQUI COLLECTI
3. three levels of data input are provided: the first level will include all essential data such as ancient name of the discovery place, modern name of the discovery place, ancient region, modern state, publication in which the text can be found and concordances with all other editions, epigraphic text, date, registration editor, state of elaboration of the text, identification number, date of registration. The second level will register data concerning the monument of the inscription: the third level data related to the content of the inscription itself.
4. participants must use programs enabling to export data in Document Type Definition (DTD) format
5. choice of the fonts must wait until the UNICODE will be adopted, and the special needs of Greek and Latin epigraphy should be presented to the UNICODE Commission by AIEGL. Besides, for Latin the use is recommended of signs LaTeX compatible
6. next to the epigraphic database a virtual images database will be created, visible on Internet using the JPEG format
7. access to the epigraphic as well as to the image database will be totally free through Internet
8. the database born under the aegis of AIEGL will be a completely new product, independent from all existing projects, and will have its own home
9. all participants to this Round Table, in their name and in name of their Institutions, are willing to participate to this database with all the data already collected, and the database will recognize the provenience of this material
10. a specific web site will be designed having three purposes: promote dialogue among the members of the Commission, provide information about is activity , collect any news about informatic projects related to epigraphy which could be already existing, in progress or to be (5).

The second phase: from Tituli Antiqui Collecti to Eagle

The promoters soon realized that in order to reach the target a series of problems had to be overcome:

1. funding: 25.000 euro were provided form the start by the Consiglio Nazionale delle Ricerche (CNR), the Commissione per le Inscriptiones Italiae of the Unione Accademica Nazionale, EDH and other Italian institutions
2. to find a computing center to assist the project: CISADU of the University of Rome La Sapienza was ready to offer assistance and to host the database when ready
3. the definition of the first level cataloguing template: the template of EDH was taken as a model, and a manual was compiled
4. the fonts problem: since no particular Unicode fonts were available for this purpose, it was decided to adopt the already established 'Epigraph' character, created by the University of Bari, which provides all diacritic signs needed by the transcription system Krummey-Panciera



5. to experiment with data transfer and input: several groups of inscriptions coming from Rome, Bari, Genoa and Trieste were chosen to be transferred, and new inscriptions were put in: both operations went smoothly
6. to find a program for data archiving and processing: still uncertain was the choice between DB2, the program adopted by EDH, and MySQL. (This uncertainty, though, did not last long because it was decided to adopt the open source MySQL).

In Trieste and Aquileia in November 2003 the International congress 'Dall'orto lapidario a Internet: un nuovo strumento per lo studio della storia antica' was held, organized by prof. Claudio Zaccaria of the Epigraphic Laboratory of the University of Trieste, and on that occasion the 2nd round table of the Commission 'Epigraphie et Informatique' of AIEGL was held.

The Commission aimed at assessing the state of the art of the database project which started in 1999. After an extensive debate the following decisions were taken:

- the experimental phase is regarded as completed
- the database will not be unique but will be the outcome of a federation of many databases under one unique portal. Federation and portal will take the name of EAGLE, namely Electronic Archive of Greek and Latin Epigraphy
- at first the federated databases will be EDH, EDR, EDB
- the federation will follow the rules of a statute, and its governing board will be set up within 2004
- the three databases will collect epigraphies as follows: Roman provinces (except Sardinia and Sicily) EDH; Christian epigraphy EDB; Rome (except Christian epigraphies), Italian Peninsula, Sicily and Sardinia
- collaboration to the databases is open according to the established rules
- the portal will reside in Rome. Technical details of the functioning will be agreed among the federated databases
- input templates, as well as input criteria and search

templates are agreed upon and publicized in the sites of the three databases, in agreement with EAGLE
– the database is protected by the European Laws in this field

All these decisions were taken unanimously and were delivered to the Bureau of AIEGL from the Commission, which is well aware to have accomplished its task once EAGLE has been established.

However,, there are still two problems to be tackled: Greek epigraphy, especially for the oriental areas before and after the roman occupation, should enter EAGLE (the main electronic archives now existing are those from Cornell University and from the Centre for the Study of Ancient Documents of Oxford). Project PETRAE, despite its independency and its specific goals, should find a way to cooperate with EAGLE.

The Portal Eagle

Once the idea of a unique database was abandoned, the portal eagle became a reality and it is searchable at the following address http://www.eagle-eagle.it (6)

The fact must be stressed that this is a low budget project because technicians chose only low cost and freely available tools and software: MySQL 4.1 for the database, Linux Slackware 8.1 as system; a new server is going to host the portal soon and it will be the only expense sustained in 6 years by the Rome Department in La Sapienza.

The template now available allows to search simultaneously through the federated databases, and the technicians of EAGLE are trying to create scripts able to search the databases and give back results consistent and similar in content.

The experts are now busy in comparing all fields of the 3 databases with the aim to enrich the search template, and in the future more fields will be added: place of discovery, date, inscription typology.

The CD-ROM Epigraph of prof. Jory will be added as searchable source of the portal instead of being transferred in EDR.

The Portal EAGLE is of course a work in progress and we all hope that more and more databases will join the project in the future, thus enabling scholars to find in it the most exhaustive tool of research for Latin and Greek Epigraphy.

References

1. Last visited 30.04.2005
2. Last visited 30.04.2005
3. Last visited 30.04.2005
4. Epigraphica, 60, 1998, pp. 314-317
5. Epigraphica, 61, 1999, pp.311-313
6. Last visited 30.04.2005

Layers and dimensions. The representation of complex structured sources

*Matthias P. Perstling**

Since Karl Lachmann (1793-1851) stated more precisely the methods of creating a critical text-edition, it is the aim of the editorial studies to represent a text in pureness and completeness in a concise and readable way. That applies to both philological and historical editions, even though the specific focusing of the disciplines leads to slight modifications of the representation. A historical edition tries to facilitate answering historian's political, social, ecclesiastical, etc. questions. Furthermore, it provides information about the adoption of text as well as existing parallel traditions. This is the great difference to philological editions, that attribute more importance to variants of writing and thus, to the critical apparatus of variants.

Nevertheless, methodology as well as techniques of critical editing in both disciplines were and are considered to be sophisticated. As a result, hardly any changes in the techniques were made during the last 150 years. All the more saddens the circumstance that a large amount of interesting historical sources has never been edited critically or, if so the attempts collapsed. There are various reasons for this failure, like for example, that financing of such projects plays an important role or that some sources are considered to be not worth being edited, but the principal reason definitely is, that many sources are much too *complex* to be edited by the traditional methods of critical edition in a printed form.

In this context *complex* means, that multiple '*nonlinear*' layers are stratified one upon the other. In contrast, a text is *linear*, if it stands only for itself. As the ideal *linear* text, we have to imagine an autograph which was written by only one person, has no outward links, is self-contained and has not affected other texts. The less *linear*, and thus, more *complex* a text is, the more complicated is its representation.

The complexity of a historical source can take place on different levels and it is also coined by the type of text of the source. So, on the one hand we have to denominate a source as *complex*, if various traditions were produced by it or it was composed of several opuses, even though it is *the* conserved original text written by one single person. Examples for an attempt of editing a source of such type are the almost unusable critical edition of the *liber pontificalis* by Theodor Mommsen (MGH, Gesta.) or the unfinished edition of the collection of the *Capitularia Benedicti Levitae* (Schmitz, Unvollendet.). On the other hand, an autograph without any links to other texts has to be considered as *complex*, if the source is full of deletions, glosses and addendums of different scribal hands or if it has a multi-layered internal structure. This category is exemplified by historical sources like necrologies, libri memoriales or late medieval manuscripts such as tax or law books.

Methodological considerations

Has the scientific community to resign to the fact that a large amount of historical sources with a more complex internal structure will never be critically edited in a satisfying way? Or are there other possibilities to represent a source which exceeds the traditional methods with their two-dimensional systems in a printed form?

For answering these questions, we have to survey the above mentioned goal of a critical edition, namely

*Historical Information and Documentation Science, Graz University

that it ought to represent a text in pureness and completeness in a concise and readable way. For I. Kropač, to edit a source means rather to represent the information structure, the information content, the information denseness and the structure of the historical tradition of a source, in a way which should be as complete as possible and enables the users of the edition to use it in an analytic way without being obliged to see the original source. (Kropač, Theorien 306. In original terms: 'Eine Quelle zu edieren bedeutet vielmehr die Informationsstruktur, den Informationsgehalt, die Informationsdichte und die Überlieferungsstruktur einer Quelle auf eine Weise zugänglich zu machen, die so vollständig als möglich sein soll und den Benutzer einer Edition in die Lage versetzt, diese analytisch zu verwenden, ohne die eigentliche Quelle einsehen zu müssen.') In this common definition the term *information* appears for the first time. ly seen, an edition is basically nothing else than the transformation of information from one medium to another. The only problem is that a transformation always causes a loss of information. In this particular case, a loss of information is a decrease of the source's authenticity. The editor tries to reduce the loss of information to a minimum, but he/she will not succeed in doing so by editing a complex text without representing the internal structure of the source. Due to this fact, the representation of a multi-layered source has to transcend the two-dimensional system of a printed form. The only appropriate possibility of representing various layers and levels, thus, is a computer-assisted editing system which is able to illustrate the layers.

If we stated above that an edition always leads to a loss of information, then we considered only half the truth. However, though an edition means a loss of information, you get a gain of information in return. A gain of information as the edition is enlarged by the editor's knowledge about the source and his/her interpretation. But this knowledge is not linear and it has to be added to the elementary information of the edition. Therefore, any attempt of editing results in a multi-dimensional representation, even if the original source is almost linear. A critical edition thus is less *and* more complex than the original source and offers less *and* more information at the same time. The aim is to receive all the information or better the exact information the users need at a particular time for specific purposes. To facilitate the achievement of this goal, the presentation needs to be dynamic and has to offer hyper media in its genuine sense.

Recapitulating the demands on a critical edition of a complex-structured source, a system is needed, that represents the text in a computer-based dynamic way. Due to the fact that this certainly means more than posting a classically printed edition on the internet, we help ourselves with an elaborated method of digital editing: the so-called method of 'integrated computer-supported editing' (In German: Integrierte Computergestützte Edition (ICE)). In this context, 'computer-supported editing' means, that the scientific editing process relies at least partially on technical procedures, that the output can be analysed with technical methods and the implementation is based on an information system which guarantees an improved comprehension of the editor's decisions and on the edited entities. 'Integrated' stands for attaining the following goals by an iterative scientific process: transparency, traceability and the possibility of emending the edition; modularity of the system organisation; advanced availability and usability; open system technologies, which depend on international standards.

This sophisticated method has been realised in the Fontes Civitatis Ratisponensis (FCR) and has proven to be of value. For representing more complex sources, the method of 'integrated computer-supported editing' admittedly needs an enlargement. We stated, that a dynamic system demanding several modifications is required for the representation of multi-layered and multidimensional sources.

For operating a dynamic system, we have to keep in mind that a database is essential. To handle easily the information stored in a database, the data has to be recorded in a well-structured form. Of prime importance thereby is, that the structure or model of the data is to reproduce the source without considering the edition's appearance. In other words, there must be a strict separation of content and form. In the same way various layers or variants of a complex source were treated. They have to be distinguished and marked separately.

The dynamic part of the system is to be such, that the users may compose their 'own' editions. So the users' specific demands on the source shall be recognisable and are to be converted into a useful representation. This postulation requires a deliberate user interface and different possibilities of access to the system. In this connection, we have to consider some facts. First of all, it must be able to cite the edited text, even if the version is generated in a dynamic way. Fur-

thermore, the content of the source must be precisely distinguishable from the editor's knowledge about it.

This enlargement of the method of 'integrated computer-supported editing' for complex sources may appear a little bit and not precisely defined, but this was intended (with full awareness of its inaccuracy). P. Sahle stated, that different sources required different methods of editing (Sahle, Edition.) and the author totally agrees with that. Due to this fact, *one* elaborated method for all kinds of complex sources seems useless. The more specialised a method is, the more exceptions and modifications are needed for the various examples. The above mentioned method is only to set up a frame in which the editions of many different kinds of complex sources can be realised using these basic requirements.

A case study: The *Marchfutterurbar* of 1414/1426

As an example for an edition of a complex source based on this method, we want to refer to the representation of the 'steirisch-landesfürstliche *Marchfutterurbar* von 1414/1426' (MFU), a manuscript of the Styrian Regional Archive in Graz, Austria. This late medieval tax book of the Styrian duke's income for his cavalry was used over a period of at least 12 years, in which numerous modifications were made by various scribes. The importance of this source for the regional history lies in the fact, that for the first time in history a huge number of taxable peasants were itemised by name in a large area of Styria. The circumstance, that the source contains so many deletions and addendums due to the frequent reuse of the manuscript led to its multi-layered internal structure and to the absence of an adequate edition. Only one attempt of editing it, 'succeeded' in 1910 (Dopsch, Gesamturbare 311-590.), but in an unsatisfaying way; it is in table form without regarding its prosopographical relevance which is due to omitting the peasants' names.

The goal of a satisfactory edition of the *Marchfutterurbar* is to represent each current state of the source at each particular moment and, thus, rendering all persons named in an obvious way connected with the respective tax. Associated with this ambition, the users must have the possibility to isolate each specific layer of the text. On the other hand, the users want to receive the exact information needed for their explicit purposes. That requires a range of different options to access the system.

First of all, the users shall have the possibility to view the facsimiles of the source (see example Figure 1.). Even the best critical edition cannot visualise the source and consequently the loss of information regarding the decrease of the source's authenticity can be reduced. Furthermore, the advantage of hyper media in a computer-based system should be utilised. So, the users can digitally browse through the manuscript and switch to the edition at every position they want to. Additionally, a view is generated, in which the image and the edition are displayed side by side and can be compared easily. This view is the default setting of the system.

The most common access to a critical edition leads the users to the text itself, though the whole text is seldom read from the very first beginning to the end. Particularly tax books like the *Marchfutterurbar* were consulted to receive specific information. To facilitate this retrieval process, several implements are installed

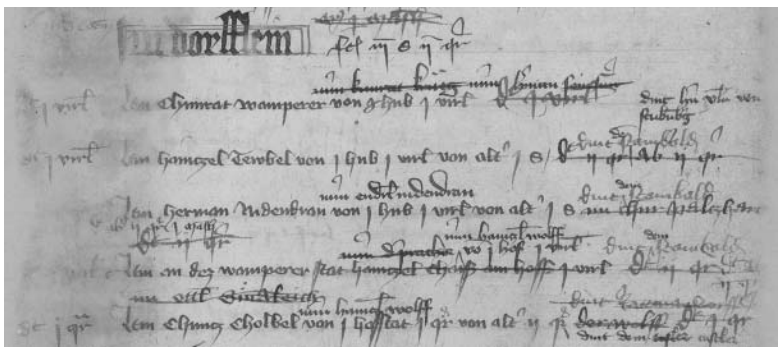


Figure 1. Text passage of the MFU, fol. 17r

in the system like the indexes or catalogues of persons and places and the glossary. In the system, simple full-text search as well as advanced search possibilities are integrated. Thus, the information required by users of all kinds of disciplines can be found. But in what way is it represented? The traditional form of the edited text with its huge critical apparatus which is almost unusable for the *Marchfutterurbar* has to be noted first (see the example in Figure 2). As mentioned above, the problem is its multi-layered internal structure due to the numerous deletions and addendums, and this is why much information vanishes in the apparatus of variants. This traditional form can be represented in two different ways, either considering the metaphor of pages or the internal structure of the 'Marchfutterurbar', i.e. organised by parishes and villages.

Indeed, the proper innovation of the computer-based edition of the *Marchfutterurbar* with the representation of its separated layers, is the more convenient access to the text. In this source, five layers can be found, the first denominating the state before 1414 (based on a vanished draft); the current state of 1414 with the particular taxes; the state between the years of 1414 and 1426, a period within which several deletions and addendums were made; the current state of 1426 and at last, the state after the terminating modifications. These five layers can be displayed either singularly or comparatively, one below the other (see the example in Figure 3). So, the users are able to apprehend easily the various states and the modifications on the text. Continuative palaeographical researches can also be made with no difficulty, as the various scribal hands can be highlighted and compared with the facsimile.

3.16		Im Dorfflein[l] *facit 3 s 2 qr.*[a]	
3.16.1	*Dedit 1 virl.*[b]	Item Chunrat Wamperer[c] *nun Herman*[d] von ½ hub 1 virl.[e]	*Dint hern Vlr(eich) von Stubenberg.*[f]
3.16.2	*Dedit 1 virl.*[g]	Item Hainczel Tewbel von 1 hub 1 virl von alter 1 s.[h]	*Dint dem*[i] *Stainbald.*[j]
3.16.3	*[Dedit ...] qr [...] Dedit aber 2 qr 1 mash.*[k]	Item Herman Nidendran *nun Enderl Nidendran*[l] von 1 hub 1 virl von alter 1 s.[m]	*Dint dem*[n] *Stainbald.*[o]
3.16.4	*Dedit [...] virl c.*[p]	Item an dez Wamperer stat *nun Hainczel Wolff*[q] *von 1 hof 1 virl*[r].[s]	*Dint dem*[t] *Stainbald.*[u]
3.16.5	*Dedit 1 qr.*[v]	Item Chuncz Cholbel *nun Hainczel Wolff*[w] von 1 hofstat 1 qr von alter 2 qr.[x]	*Dint dem*[y] *Tastler.*[z]

3.16 [a] *Von späterer Hand (H7) nachgetragen.* [b] *Von späterer Hand (H4) nachgetragen.* [c] *Danach über der Zeile nachgetragen von späterer Hand (H3) und von späterer Hand (H5) gestrichen:* nun Kunrat Krüg. [d] *Über der Zeile von späterer Hand (H5) nachgetragen. Danach über der Zeile von späterer Hand (H5) nachgetragen und von späterer Hand (H6) gestrichen:* sein sun. [e] *Danach der Nachtrag von späterer Hand (H2), der von späterer Hand (H4) wieder gestrichen wurde:* Dedit 1 virl. [f] *Von späterer Hand (H5) nachgetragen..* [g] *Von späterer Hand (H4) nachgetragen.* [h] *Danach von späterer Hand (H2) nachgetragen und von späterer Hand (H4) gestrichen:* Dedit 2 qr. Aber 2 qr. [i] *Über der Zeile von späterer Hand (H5) nachgetragen.* [j] *Von späterer Hand (H2) nachgetragen.* [k] *Von späterer Hand (H4) nachgetragen.* [l] *Über der Zeile von späterer Hand (H4) nachgetragen.* [m] *Danach von späterer Hand (H2) nachgetragen und von späterer Hand (H4) gestrichen:* Nu Chunrat Palczhart. Dedit 2 qr. [n] *Über der Zeile von späterer Hand (H5) nachgetragen.* [o] *Von späterer Hand (H2) nachgetragen.* [p] *Von späterer Hand (H4) nachgetragen.* [q] *Über der Zeile von späterer Hand (H5) nachgetragen, statt dessen der Nachtrag von Hand (H4) über der Zeile gestrichen:* nun der Pracher. [r] *Über der Zeile von späterer Hand (H4) nachgetragen, statt dessen gestrichen:* Hainczel Chaiss ain hoff 1 virl. [s] *Danach der Nachtrag von späterer Hand (H2), der von späterer Hand (H4) wieder gestrichen wurde:* Dedit 2 qr. Dedit aber 2 qr. *Danach von späterer Hand (H3) nachgetragen und von späterer Hand (H4) gestrichen:* Nun Ottel Sindleich. [t] *Über der Zeile von späterer Hand (H5) nachgetragen.* [u] *Von späterer Hand (H2) nachgetragen.* [v] *Von späterer Hand (H4) nachgetragen.* [w] Über der Zeile von späterer Hand (H5) nachgetragen. [x] *Danach von späterer Hand (H2) nachgetragen und von späterer Hand (H4) gestrichen:* Der Wolff. Dedit 1 qr. [y] *Danach von derselben Hand (H5) gestrichen* Tasler. [z] *Unter der Zeile geschrieben, statt dessen der Eintrag von früherer Hand (H2) über der Zeile gestrichen:* Dint Rattmanstorffer.

[l] Dörfl (R), KG Ponigl, OG Thannhausen, GB u. BH Weiz.

Figure 2. Traditional edition of the text passage above (MFU, fol. 17r); the underlines stand for rubrications

3.16a		Im Dorfflein[1]	
3.16c		Im Dorfflein[1] facit 3 s 2 qr.	
3.16.1a		Item Chunrat Wamperer von ½ hub 1 virl.	
3.16.1b		Item Chunrat Wamperer von ½ hub 1 virl. Dedit 1 virl.	
3.16.1b'		Item Chunrat Wamperer nun Kunrat Krüg von ½ hub 1 virl. Dedit 1 virl.	
3.16.1b''	Dedit 1 virl.	Item Chunrat Wamperer nun Herman sein sun von ½ hub 1 virl.	Dint hern Vlr(eich) von Stubenberg.
3.16.1c	Dedit 1 virl.	Item Chunrat Wamperer nun Herman von ½ hub 1 virl.	Dint hern Vlr(eich) von Stubenberg.
3.16.2a		Item Hainczel Tewbel von 1 hub 1 virl von alter 1 s.	
3.16.2b		Item Hainczel Tewbel von 1 hub 1 virl von alter 1 s. Dedit 2 qr. Aber 2 qr.	Dint Stainbald.
3.16.2c	Dedit 1 virl.	Item Hainczel Tewbel von 1 hub 1 virl von alter 1 s.	Dint dem Stainbald.
3.16.3a		Item Herman Nidendran von 1 hub 1 virl von alter 1 s.	
3.16.3b		Item Herman Nidendran von 1 hub 1 virl von alter 1 s. Nu Chunrat Palczhart. Dedit 2 qr.	Dint Stainbald.
3.16.3b'	[Dedit ...] qr [...] Dedit aber 2 qr 1 mash.	Item Herman Nidendran nun Enderl Nidendran von 1 hub 1 virl von alter 1 s.	Dint Stainbald.
3.16.3c	[Dedit ...] qr [...] Dedit aber 2 qr 1 mash.	Item Herman Nidendran nun Enderl Nidendran von 1 hub 1 virl von alter 1 s.	Dint dem Stainbald.
3.16.4a		Item an dez Wamperer stat Hainczel Chaiss ain hoff 1 virl.	
3.16.4b		Item an dez Wamperer stat Hainczel Chaiss ain hoff 1 virl. Dedit 2 qr. Dedit aber 2 qr.	Dint Stainbald.
3.16.4b'		Item an dez Wamperer stat Hainczel Chaiss ain hoff 1 virl. Dedit 2 qr. Dedit aber 2 qr. Nun Ottel Sindleich.	Dint Stainbald.
3.16.4b''	Dedit [...] virl c.	Item an dez Wamperer stat nun der Pracher von 1 hof 1 virl. Dedit 2 qr. Dedit aber 2 qr.	Dint Stainbald.
3.16.4c	Dedit [...] virl c.	Item an dez Wamperer stat nun Hainczel Wolff von 1 hof 1 virl.	Dint dem Stainbald.
3.16.5a		Item Chuncz Cholbel von 1 hofstat 1 qr von alter 2 qr.	
3.16.5b		Item Chuncz Cholbel von 1 hofstat 1 qr von alter 2 qr. Der Wolff. Dedit 1 qr.	Dint Rattmanstorffer.
3.16.5c	Dedit 1 qr.	Item Chuncz Cholbel nun Hainczel Wolff von 1 hofstat 1 qr von alter 2 qr.	Dint dem[a] Tastler.

[a] *Danach gestrichen* Tasler.

[1] Dörfl (R), KG Ponigl, OG Thannhausen, GB u. BH Weiz.

Figure 3. Representation of the separated layers of the aforementioned text (MFU, fol. 17r); the underlines stand for rubrications, variants 'a', 'b', etc. etc., for the various layers

Another possibility to break through the linearity and two-dimensionality of a printed edition and consequently creating a new way of access to it, is the option of displaying different entries related to one specific topic. So, the users can for example, retrieve all peasants of one particular lord of manor and display the entries simultaneously. That facilitates the work with the source and comes certainly to the historian's aid, whose interests are in regional history, in genealogy or prosopography.

For the users who search for information on a specific geographical region, the system proposes an additional access. On a digital map all villages mentioned in the *Marchfutterurbar* are plotted and linked to the entry in the edition. Due to this fact coherences between two villages can be pointed out more easily, even if their entries in the source are separated by a dozen pages of the manuscript. For example, it would be more precise and easier to locate a larger area where crop failure happened, if this was mentioned in several passages of the source in different places. This might be of great relevance for economic historians, but there are also other representations of the source especially for them. For the users who are only interested in the amount of the taxes, the system offers a representation in table form with the possibility to analyse and evaluate it statistically.

Another access is provided for specialists in palaeography or codicology. They can enter the system via the palaeographical or codicological comments, where, for instance, the different scribal hands, the water marks or the collation formula of the manuscript are described. The examples are of course always supplied with links to both the facsimiles and the edited text.

All these mentioned ways of accessing the computer-based edition of the *Marchfutterurbar* result in a sys-

tem of a dynamic, multi-dimensional representation of the source. To handle all these various possibilities of access, an advanced user-interface is needed. This interface accesses the data embedded in the system using one of the above-mentioned possibilities. In the system the data exist in different conditions. On the one hand, the transcribed raw text is marked with the Extensible Markup Language (XML). Thus, the various scribal hands and layers are also represented by means of a Document Type Definition (DTD) that is based on the internal structure of the source. To generate a dynamic edition of the XML data and to control the several layers, technical solutions are planned that either are based on SAX (the Simple API for XML) or the DOM (the Document Object Model) or convert the data to a native XML database operated by the XML Query Language ('XQuery').

For a simple text output, this would be a feasible way, but our intentions transcend these customs of representation. We already pointed out that for analysing and evaluating the source statistically a new option is required. To provide for this aims, we use databases that allow processes of aggregation, the use of ontologies and other features of analysis and that can handle the structure of complex sources. For exactly these intentions, the software package Κλειω was created and operates in this field with wide satisfaction. Thus, we are working with a hybrid form of data. XML data are read in the Κλειω databases, where they are analysed, stored and enriched, and finally the output is generated as modified XML data.

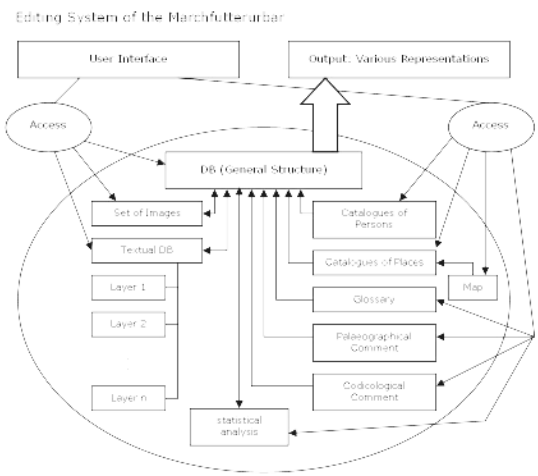


Figure 4. Scheme of the whole system

Summary

Our goal is a critical edition of a special type of source, which has not yet been edited due to the complexity and resulting difficulties of the representation with its multi-layered internal structure. The only expedient way of representing the information structure, the information content, the information denseness and the *context* of the historical tradition of the *Marchfutterurbar*, which should be implied by a critical edition, is to create a dynamic computer-supported editing system. This system offers various ways of accessing the source, allows the users to distinguish easily between the different layers and enables them to display a representation of the *Marchfutterurbar* with regard to their own demands. The constraints of a critical edition in a printed form with its linearity can be overcome and the multi-dimensionality provides an additional value of information.

References

Heidrun Boshof, Die Fontes Civitatis Ratisponensis: Geschichtsquellen der Reichstadt Regensburg online. In: Klaus van Eickels / Ruth Weichselbaumer / Ingrid Bennewitz (Eds.), Mediävistik und Neue Medien (Ostfildern 2004), p. 279-294.

Susanne Botzem / Ingo H. Kropač, Integrated Computer Supported Editing, Approaches and Strategies. In: Historical Social Research 16 / 4 (1991), p. 106-115.

Dino Buzzetti, Ambiguità diacritica e markup. Note sull'edizione critica digitale (2000) [Internet: http://dobc.unipv.it/dipslamm/pubtel/Atti2000/dino_buzzetti.htm, 2005-04-18].

Document Object Model (DOM) Level 3 Core Specification. Version 1. W3C Recommendation 07 April 2004 [Internet: http://www.w3.org/TR/DOM-Level-3-Core/, 2005-04-18].

[Dopsch, Gesamturbare.]
Alfons Dopsch, (Ed., with Alfred Mell), Die Landesfürstlichen Gesamturbare der Steiermark aus dem Mittelalter (= Österreichische Urbare, I. Abt., Bd.2, Wien – Leipzig 1910), p. 311-590.

Extensible Markup Language (XML) 1.1, W3C Recommendation 04 February 2004, edited in place 15 April 2004 [Internet: http://www.w3.org/TR/xml11/, 2005-04-18].

Horst Fuhrmann, Über Ziel und Aussehen von Texteditionen. In: Mittelalterliche Textüberlieferungen und ihre kritische Aufarbeitung. Beiträge der Mo-

numenta Germaniae Historica zum 31. Deutschen Historikertag Mannheim 1976 (München 1976), p. 12-27.
Fontes Civitatis Ratisponensis (FCR) [Internet: http://www.fcr-online.com, 2005-04-20].
Claus Huitfeldt, Electronic Textual Editing: Philosophy Case Study [Internet: http://www.tei-c.org/Activities/ ETE/Preview/huitfeldt.xml, 2005-04-20].
Integrierte Computergestützte Edition (ICE) [Internet: http://www.fcr-online.com/ice, 2005-04-20].
Ingo H. Kropač, Ad fontes. Von Wesen und Bedeutung der Integrierten Maschinellen Edition. In: Herwig Ebner / Horst Haselsteiner / Ingeborg Wiesflecker-Friedhuber (Eds.), Geschichtsforschung in Graz. Festschrift zum 125-Jahr-Jubiläum des Instituts für Geschichte der Karl-Franzens-Universität Graz (Graz 1990), p. 465-482.
[Kropač, Theorien.]
Ingo H. Kropač, Theorien, Methoden und Strategien für multimediale Archive und Editionen. In: Klaus van Eickels / Ruth Weichselbaumer / Ingrid Bennewitz (Eds.), Mediävistik und Neue Medien (Ostfildern 2004), p. 295-316.
Ingo H. Kropač / Susanne Botzem, As You Like It. Archiving, Editing and Analysing Medieval Documents. In: Josef Smets (Ed.), Histoire et Informatique. Ve Congrès 'History & Computing', 4. – 7. 9. 1990 à Montpellier (Montpellier 1992), p. 301-313.
Ingo H. Kropač / Susanne Botzem / Henriette Kurschel, Das ICE-Projekt. In: Informatik Forum 8/4 (1995), p. 159-164.
Ingo H. Kropač/ Susanne Kropač, Prolegomena zu einer städtischen Diplomatik des Spätmittelalters, das Beispiel Regensburg. In: Walter Prevenier / Thèrés de Hemptinne (Eds.), La diplomatique urbaine en Europe au moyen âge. Actes du congrès de la Commission internationale de Diplomatique, Gand, 25-29 août 1998 (= Studies in Urban Social, Economic and Political History of Medieval and Early Modern Low Countries 9. Louvain / Apeldoorn 2000), p. 229-265.
[MFU] = Steirisch-landesfürstliches Marchfutterurbar von 1414/1426.
Steiermärkisches Landesarchiv (StLA), Laa. A. Antiquum II, Stockurbar 43/64.
[MGH Gesta.]
Monumenta Germaniae Historica Gesta Pontificum Romanorum 1,1 (Berlin 1898).
Malte Rehbein, Die dynamische digitale Textedition: Ein Modell. In: Hans-Heinrich Ebeling et al. (Eds.), Vom digitalen Archiv zur digitalen Edition. Begleitheft zur CD-ROM (Göttingen 1998), p. 5-22.
Malte Rehbein, Edition als Softwareproblem: Die 'dynamische Textedition'. In: Concilium medii aevi 1 (1998), p. 1-15.
Malte Rehbein, Die digitale Textedition. In: Hans-Heinrich Ebeling / Manfred Thaller (Eds.), Digitale Archive. Die Erschließung und Digitalisierung des Stadtarchivs Duderstadt (Göttingen 1999), p. 103-124.
[Sahle, Edition.]
Patrick Sahle, Digitale Edition (Historischer Quellen) – Einige Thesen (1997) [Internet: http://www.uni-koeln.de/~ahz26/ dateien/thesen.htm, 2005-05-13].
Patrick Sahle, Digitale Editionstechniken und historische Quellen. In: Stuard Jenks / Stephanie Marra (Eds.), Internet-Handbuch Geschichte (= utb für Wissenschaft 2255, Köln / Weimar / Wien 2001), p. 153-166.
[Schmitz, Unvollendet.]
Gerhard Schmitz, 'Unvollendet' – 'Eingestampft' – 'Kassiert'. Nie Erschienenes und Mißglücktes. In: Zur Geschichte und Arbeit der Monumenta Germaniae Historica. Ausstellung anlässlich des 41. Deutschen Historikertages München, 17.-20. September 1996: Katalog (München 1996), p. 64-73.
Manfred Thaller, Ungefähre Exaktheit. Theoretische Grundlagen und praktische Möglichkeiten einer Formulierung historischer Quellen als Produkte 'unscharfer' Systeme. In: Herta Nagl-Docekal / Franz Wimmer (Eds.), Neue Ansätze in der Geschichtswissenschaft (= Conceptus-Studien, Wien 1984), p. 77-100.
Manfred Thaller, Secundum Manus. Zur Datenverarbeitung mehrschichtiger Editionen. In: Reinhard Härtel (Ed.), Geschichte und ihre Quellen. Festschrift für Friedrich Hausmann zum 70. Geburtstag (Graz 1987), p. 629-637.
Manfred Thaller, Datenbasen als Editionsformen. In: Anton Schwob / Karin Kranich-Hofbauer / Diethard Suntinger (Eds.), Historische Edition und Computer. Möglichkeiten und Probleme interdisziplinärer Textverarbeitung und Textbearbeitung (Graz 1989), p. 215-241.
Manfred Thaller, Κλειω 3.1.1. Ein Datenbanksystem

(= Halbgraue Reihe zur Historischen Fachinformation B 1, St. Katharinen 1991).
Manfred Thaller, Κλειω. A Database System (= Halbgraue Reihe zur Historischen Fachinformation B 11, St. Katharinen 1993).
Manfred Thaller, Vom verschwindenden Unterschied zwischen Datenbanken und Texten: Konsequenzen neuerer www-Technologie am Beispiel von museumsnahen Datenbanken (2000) [Internet: http://www.museumtheuern.de/edvtage/g_mat/g12_thal.pdf, 2005-04-20].
Manfred Thaller, Texts, Databases, Κλειω: a Note on the Architecture of Computer Systems for the Humanities. In: Dino Buzzetti / Giuliano Pancaldi / Harold Short, Augmenting Comprehension. Digital Tools and the History of Ideas (= Office for Humanities Communication Publication 17, London 2004), p. 49–76.
Gunter Vasold, Edition à la carte? Usability, Interfacing und Datenmigration für webbasierte Editionssysteme. In: Klaus van Eickels / Ruth Weichselbaumer / Ingrid Bennewitz (Eds.), Mediävistik und Neue Medien (Ostfildern 2004), p. 261-278.
Virtual Library Geschichte, (Digitale) Editionstechnik / Scholarly Digital Editing (2002) [Internet: http://www.uni-koeln.de/~ahz26/vl/editech.htm, 2005-04-18].
XML Query (XQuery) [Internet: http://www.w3.org/XML/Query, 2005-04-18].

Transnational histories in Roshini Kempadoo's *ghosting*. (Cyber)Race identities

*Sheila Petty**

For digital artists of the black diaspora, 'blackness in particular, is the anti-avatar of digital life,' and simply being accounted for in the histories arising from new media is an extension of other, earlier struggles to establish the legitimacy of their own historical experiences (Nelson 1). As a result of this 'constant battle to exist,' these artists forge new ways of conceiving and recouping histories obscured and, indeed lost, in the continued struggle for identity in postcolonial contexts (Oguibe 35).

In this paper, I intend to explore how Roshini Kempadoo's black diasporic new media artwork *Ghosting* meets this challenge by dismantling western linear constructs of history to reconstitute it as a transnational flow of impulses contributing to complex contemporary Caribbean identities. In particular, I will consider how the recombinant narrative structure exploits digital compositing, archival materials and narration to challenge fixed notions of race and globalization. I will thus demonstrate that Kempadoo offers a compelling example of how personal and global histories collide, fragment and transcend western imperatives to provide an engagement with digital aesthetics that is specific to black diasporic digital imaginings.

A question of criteria

Generally, a discussion such as the one proposed here begins with an apologia of sorts that argues for the continued importance of race and race discourses in resistance to euro-based new media theory and that the body, and with it race, have become irrelevant in the face of computer-based technologies.[1] Instead, I propose a very different approach and take the position that such post-human new media theory is merely one way of explicating how we experience digital technologies from a distinctive cultural perspective. For example, in the case of the black diaspora, an already existing plethora of theory not only challenges the universalist tendencies of current euro-based post-human theoretical constructs, but also lays the groundwork for developing entirely different criteria by which to measure how black experience is expressed within cyberspace.

For many in the Caribbean black diaspora, globalization is far from a recent phenomenon. Forged in the crucible of slavery by Spanish, Dutch, Portuguese, English, and French colonial powers, the black cultures of the Caribbean suffered dehumanization as slaves' bodies became the labor that fueled the Industrial Revolution. In addition, many Caribbean nations, in an attempt to sustain plantation economies after emancipation, sought out waves of laborers from India, China and other countries and frequently oppressed them as well. As H. Adlai Murdoch argues, 'it is the plurality, the discontinuity, and the dispersal inherent in these historical and cultural elements which go together to make up the unique quality of the Caribbean heritage' (4). Given this context, it is clear that totalizing euro-based new media theoretical constructs of subjectivity seeking to quantify a reductive experience of computer culture would be at odds with Caribbean identities rooted in an environment where 'factors of racial and cultural pluralism,' in concert with competing global histories, encourage 'specificity and, at the same time, inhibit the affirmation of unity and identity (4-5). In methodological terms, new criteria for measuring a polyvalent experience of cyberspace that is culturally specific and responsive to the existence of multiple histories must be sought to explicate black Caribbean experience.

*University of Regina

Where then can such criteria be uncovered? The Caribbean, like much of the black diaspora, has a rich resource of critical thought that can be adapted to use in the area of new media, just as euro-based new media theory draws on earlier works by authorities such as Jean Baudrillard or Michel Foucault. For example, the role played by the intersection of multiple histories and cultures has long been a major focus of Caribbean theorists such as Edouard Glissant. He argues that one of the misleading consequences of adopting a eurocentric vision of history 'is the insistence on fixing the object of scrutiny in static time, thereby removing the tangled nature of lived experience and promoting the idea of uncontaminated survival' (14). Instead, Glissant abandons notions of fixed identity 'in absolute terms' by suggesting that crosscurrents of manifold cultural components act as 'active agents of synthesis' through an endless process of mixing or creolization (16). In turn, the 'converging' cultures of the Caribbean render impotent 'the linear, hierarchical vision of a single History' so typical of eurocentric constructs (66). By doing so, Glissant draws a distinction between 'the 'totalizing' impulse of a transcendental History (with a capital *H*) and the true shapelessness of historical diversity' (Dash xxix in Glissant).

It is this gap between fixed and mutable histories and identities that British-based new media artist Roshini Kempadoo explores in her digital installation art work, *Ghosting*.[2] As an artist, Kempadoo's personal oeuvre draws on her own transnational identity, which brings together British, Guyanese, Afro-Caribbean and Indo-Caribbean inflections.[3] Reflecting these influences through its focus on 'the presence of the forced labourer, indentured individual, [and] ex-slave in the physical location of the plantation' in Trinidad 'around the period of 1838-1948,' *Ghosting* becomes a meditation on the complicated intersections between occulted personal histories and the global historical confluences that shape their realities and resistance to oppression.[4] Furthermore, by foregrounding the personal narratives of the characters in the artwork as lived experiences, *Ghosting* offers a powerful counter-history that transcends the bias of western history by uncovering the absent 'black subject' and returning her/him to the center.[5]

To understand just how Kempadoo uses the possibilities afforded by digital technology as central to her narrative goal, it is worthwhile considering how interactivity is created and sustained for the artwork's user. As the user enters the installation space, she/he encounters a console facing a screen. On the console is a warri board, comprised of rough wood. Along its upper surface, the warri board features five carved pits, four of which contain heavy stones that have smooth, worn surfaces. The pits contain interactive on/off switchers that are sensitive to the weight of the stones, and as the user moves stones from pit to pit, or removes two or more stones from the board, specific story strands are triggered and projected on the screen. The user is then free to experience a portion of, or the entire story strand, depending on whether she/he leaves the stones in place or reconfigures their order.

The warri board becomes a very powerful, tactile metaphor for the various histories that unfold through *Ghosting*'s digital narrative. As an object, the warri board has a long, trans-global history. Played across a wide swath of west Africa, Egypt, the Middle East, Asia and the Caribbean, the warri board is a game of strategy where two players move tokens with the goal of capturing each other's markers.[6] It is generally accepted that the game traveled to the Caribbean with African slaves and continues to be played in a contemporary context. Because of this history, the warri board becomes a portal to past and present by virtue of the connection of this artifact to slavery as well as its digital rearticulation by Kempadoo's computer technology. Furthermore, the absence of the ubiquitous computer mouse in favor of the adaptation of older African-inflected methodology, presents a view of technological advancement routed through an historical path quite different from eurocentric views of digital technology. Far from experiencing the disembodiment of post-human informational strategies, the tactile shape of the warri board and the physical movement of the stones brings the physicality of the user to the foreground: in the histories of Kempadoo's narrative, the body is the site of cultural discourse and labor is one of the ways in which that discourse is manifested.

Close analysis of a digital artwork is always fraught with difficulty as the demands of formal academic writing require a level of linear organization that can be considerably different from the way in which the user experiences the artwork. In the case of *Ghosting*, the narrative frame of the piece is comprised of six story strands, all of which possess both a linear and a non-linear structure.[7] As individual pieces, each has a distinct beginning point and an ending, but as a whole, the story strands possess interconnections and con-

texts that can only be uncovered through the user's exploration. As a result, because the user chooses the sequence in which they engage with the various histories and the duration of time spent on each, the user controls how much detail is gleaned from the work. Furthermore, because different story strands are rooted in a variety of periods of time from Emancipation to 1948, the user's experience of Trinidadian multiple histories is multi-linear and fragmented, requiring active participation on the user's part in order to maintain perspective. In choosing this particular narrative formulation, Kempadoo makes ideological use of digital data retrieval to reflect a distinctly Caribbean transnational view of nation in which 'an infinite wandering across cultures' creates an environment 'where triumphs are momentary and where adaptation and *métissage* (creolization) are the prevailing forces' (Dash xxviii). Like the cultural influences of Kempadoo's characters, the histories that brought them to Trinidad are recombinant and resistant to encapsulation in western linearity.

Personal histories, transnational flows

As the central touchstone of *Ghosting*, Trinidad is intended by Kempadoo to represent a transnational, trans-Caribbean imaginary. This is especially evident in the story strand that provides background information on the character of Elsie, who began life as a plantation slave. The sequence is framed as the remembrance of Aunt Ruth, a 'medicine woman normally associated with any plantation,' a device that serves to place the events of Elsie's life in the past, while providing a narrative center to the artwork.[8] Kempadoo's transnational project is laid bare when Aunt Ruth begins her recitation by observing, 'Well, I tell you, hoist up any one rock on dis island, it's the same plantation story,' whether it occurs on Trinidad, Barbados or Martinique. The statement, with its emphasis on slavery's broad impact, serves to remove Elsie's story from that of a localized personal phenomenon to one that occurs within the auspices of a significant flow of historical impetuses. Hence Elsie not only lives history but also is history in her own right.

As Aunt Ruth's remembrance unfolds, an oral image of Elsie emerges as a woman haunted by the dehumanizing effects of slavery. Ruth's image of 'the bleached, black backs' of the field hands becoming 'distorted' under 'the limbo pole of slavery as Elsie, trapped by the 'rituals of power' inside the plantation house, serves to personalize the trauma internalized by survivors of colonialism, which in turn, will become part of slavery's legacy. Kempadoo's visual images further support this contention through the use of digital compositing to layer images from the past with the present, or create imaginary landscapes by seamlessly combining several different locales into one. For example, as the narration described above unfolds, an image of a staircase appears on the right side of the screen. On the left, another staircase fades in, clearly from a different locale. However, the two are digitally matched so that they appear to flow into one another, creating a single environment. Finally, the haunted face of a black woman is superimposed over this background, seemingly suspended between the two stairways. The strong diagonal presence of the staircases and the emotive expression of the woman's face, suggest that she is trapped between her humanity and the labor that dehumanizes her.

Elsie's role as living history is further developed in the story strand dedicated to the contents of the will of Sampson de Boissiere, a white French Creole planter. As a female voice reads the will, it acknowledges that Elsie is de Boissiere's slave, and formally recognizes her daughters, Marie Louise and Victoria May, as her own. Although de Boissiere grants Elsie her freedom through this document, it nevertheless draws attention to the fact that she is a possession to be dispensed of by his estate, much the same as his house in Greenwich, which he gives to his white wife, Isabel, or the 700 acres he passes on to Jean Baptiste Phillippe in order to pay off his debts. It also places Elsie at the center of the crosscurrents of the slave trade: as a white French Creole planter, de Boissiere is representative of a period of Trinidadian colonial history beginning in 1783, when the island was held by the Spanish. At that time, the land was largely undeveloped, and in an attempt to attract planters and slaves to extract agricultural wealth from it, the Spanish encouraged French immigration from 'Grenada, Guadeloupe, St. Lucia and St. Vincent' by offering land grants based, in part, on the number of slaves an immigrant brought with them (Millette 12).[9] Kempadoo ties de Boissiere to this history through the list of his properties which include Bellside, an abandoned cocoa plantation on Grenada. In addition, along with the white French immigrants, 'free coloureds' also settled in Trinidad during this period, a fact alluded to by the mention of Jean Baptiste Phillippe in de Boissiere's will, who is later revealed in another story strand to be a leader in the free colored community (15).[10] One of the unique

aspects of this arrangement was that 'legal sanction was given to non-white planters as a property owning class,' allowing them 'security and scope' which 'was unrivalled' in the Caribbean context (17).

By tying Elsie into this historical context, Kempadoo exposes the differing stratified layers of Trinidadian history: as chattel to be dispensed of however generously and warmly by de Boissiere, she is still placed in an objectified class beneath Jean Baptiste Phillippe despite their shared origin in slavery, for not only is he black, but he is also of sufficient social position to own slaves in his own right and enter into business arrangements with a white Creole planter, an agency that is denied Elsie. Kempadoo renders this relationship visual when she couples the section of the will that grants Elsie her freedom with a full-screen shot of an archival plantation document, including a map of the property held and notations of its assets. As de Boissiere awards Elsie a small house in Laventille, its contents and all of his clothes, three small windows open vertically on the left side of the image, revealing a tennis court and a garden. These photographs connect the plantation to contemporary space, suggesting that under their reconfiguration, the history of the plantation and Elsie's own life still persist, no matter what transformation has taken place.

Personal histories, discourses of nation

Nation, in the Caribbean, is described by Glissant as 'an intellectual dream, lived at the same time in an unconscious way,' but, because of multiple, shifting histories, it seems to 'free' its peoples 'from the intolerable alternative of the need for nationalism,' introducing instead a 'cross-cultural process that modifies but does not undermine the latter' (139). However, if as Glissant contends, nation in the Caribbean is expressed 'as a multiple series of relationships,' there is also a very real possibility of fragmentation as nations and cultures are split along the tectonic fractures of histories that collide and fail to recognize the validity of one another (139). Kempadoo alludes to this state in a story strand that brings together the following characters: Jonas Mohammed Bath, is leader of the Mandingoes, ex-slaves who wish to return to Africa; Jean Baptiste Phillippe is planter and leader of the free coloreds and a practicing doctor; Ram, is a formerly indentured Indo-Trinidadian laborer who now lives with Elsie.[11] The conversation is set in the period around 1838, during the time when Trinidad was under British colonial rule, and demonstrates how the competing interests, forged in interconnected yet separate historical journeys, both serve to divide and join the men as they discuss various petitions aimed at enshrining their rights.

As the conversation opens, Bath shares the following passage from the Koran, 'have they not traveled in the land so that they should have hearts with which to understand or ears with which to hear? For surely it is not the eyes that are blind, but blind are the hearts which are in the breasts.' As well as a testimony to his deeply held faith, the passage is emblematic of Bath's recognition that slavery and the apparatus that supports it, are responsible for the great sorrow and dehumanization of his fellow African slaves. Describing himself as 'a blind man whom by the will of Allah is now able to see,' the only means he sees for rectifying this loss is to petition the British government to permit ex-slaves to return to Africa, their rightful place of origin.

Jean Baptiste Phillippe has a different view of these priorities. Under British rule, the colonial administration viewed 'the free coloureds as on the whole dangerous and unruly' and regarded their insistence on the entrenchment of the freedoms held under Spanish rule with 'an active antagonism' (Millette 109). Like Bath, the refusal of the colonial administration to deal fairly with the free coloreds necessitated turning to British parliament with petitions arguing for the validity of their position. Furthermore, because of their investment both culturally and economically in Trinidad, the free coloreds viewed themselves as committed to the island and its emerging nation, a commitment not held by Bath.[12] Hence, Phillippe takes the position that the rights of the free coloreds take precedence, a stance clearly outlined when he tells Bath that his petition for returning ex-slaves to Africa would be stronger if the petition for free colored rights passes parliament first.

Between these two Afro-Trinidadian visions of black rights, Kempadoo superimposes the views of Ram. As an Indo-Trinidadian within the historical context of the conversation, he is from an ethnic minority 'who are symbolically positioned outside the 'national self'' of Trinidad (Munasinghe 1-2). As Viranjini Munasinghe notes, 'after Emancipation in 1834, planters in Trinidad faced a dire labor situation' which they sought to deal with by securing 'a *cheap and easily manageable labor force*' (8). In order to justify their position, planters vilified Afro-Trinidadians by creating a stereotypical construct in which they were seen 'as 'free-spend-

ing, luxury loving and improvident'' in contrast to the new labor force immigrating from India who were portrayed as ''industrious, diligent, [and] self-sacrificing'' (9). Committed to ten years of labor in Trinidad 'before entitled to a free return passage to India,' Indian laborers spent the first five years of their indenture with 'severely curtailed' rights condemning them to 'the lowest status' of the Trinidadian social hierarchy because they 'were willing to do freely the work that slaves had done under coercion' (10, 9). Given that the low cost of these laborers undermined the ability of Afro-Trinidadian 'bargaining power with planters,' indentured Indians were widely 'thought of as inferior' and were thus relegated to the periphery of Trinidadian society because of 'material, social structural and cultural factors' (9, 10). It is therefore significant that Kempadoo assigns Ram the function of outsider in the conversation between competing black interests, giving him the distance to see that although Phillippe and Bath are 'fighting the same war,' they are in fact in 'different battles.' It is Ram who suggests a third view of their struggle for just legal status, arguing that the injustices faced by indentured laborers and Afro-Trinidadians could be served by simply fighting for the legitimization of all ex-slave rights. By doing so, Ram creates an equivalency between indenture and slavery that offers a bridge between histories and cultures. Although Bath and Phillippe do not adopt this position, it is a foreshadowing of the Trinidadian nation that will eventually emerge from the crosscurrents of multiple histories.

Recombinant histories, digital narratives

As *Ghosting* demonstrates, the numerous historical flows of history in the Caribbean have given rise to specific Caribbean theoretical perspectives that provide a salient framework for adaptation to new media theoretical constructs. From its articulation as a digital artwork, to the experience of shifting historical imperatives it provides the user, *Ghosting* both resists the self-serving linearity of eurocentric historical constructs and returns the Afro-Trinidadian and Indo-Trinidadian discourses to their rightful positions as subjects shaping the development of Trinidad as a nation. As Glissant observes, it is the goal of Caribbean artists to uncover 'the possibility of a collective effervescence' that rewrites the colonial project to encompass the polyphony of voices waiting to be heard (Glissant 236). Kempadoo's *Ghosting* certainly takes up this challenge and by doing so, opens new doors for expanding how digital discourses function in a cultural context.

Acknowledgements

Special thanks to D.L. McGregor for her creative insight and assistance with this essay and to R. Kempadoo for generously providing me access to her artwork. I also acknowledge the University of Regina President's Scholar program and the Social Sciences and Humanities Research Council of Canada for sponsoring this research.

Notes

1. For example, Mark B.N. Hansen argues 'by decoupling identity from any analogical relation to the visible body, on-line self-intervention effectively places everyone in the position reserved for certain raced subjects' (112). From this perspective, if we share a universal experience of body-erasure through engagement with computer technology, then race is eradicated as a factor of identity.
2. A Senior Lecturer in Digital Media at the University of East London, Kempadoo is an artist of international reputation who specializes in innovative convergence art works that explore lens-based imaging with cyber-based technology and informational networks.
3. Originally commissioned by The City Gallery (Leicester, UK) in partnership with the Peepul Centre (Leicester, UK) in 2004, *Ghosting* was one of several works in the retrospective show, *Roshini Kempadoo: works 1990-2004*, produced by OVA in association with The City Gallery. The artwork has been exhibited at international venues including Mount Saint Vincent University Art Gallery in Halifax, Canada and the Soil Digital Media Suite at Neutral Ground Artist-Run Centre in Regina, Canada.
4. R. Kempadoo, email correspondence with the author, Aug.18, 2004.
5. R. Kempadoo, email correspondence with the author, Aug.18, 2004.
6. This strategy game is known under a variety of different names including Adi (Nigeria), Omweso (Uganda) and Mancala (Egypt) to name but a few. For more information on the different variations of the game, refer to: http://gamesmuseum.uwaterloo.ca/countcap/pages/index.html
7. The monologues of the characters in the story strands were composed by dub poet Marc Mat-

thews in consultation with Kempadoo. They are written in creolise, a dialect that is itself a global product drawing influences from all the languages of slavery and post-slavery cultures including African, English, French, Spanish and Indian elements. R. Kempadoo, email correspondence with the author, Aug.18, 2004.

8. R. Kempadoo, email correspondence with the author, Aug.18, 2004.
9. As James Millette notes, 'the size of the labour problem and the thirst for land, continually interacted upon each other' (17). Whether white French Creole or Free Colored, the size of the land grant was tied directly to slave holdings: white immigrants were entitled to approximately 'thirty acres for each member of his family and half as much for each of his slaves. Blacks and people of colour, being free men and proprietors, received half of the proportion allotted to whites, the allotment to be increased if they brought slaves with them' (16).
10. To demonstrate the strength of the free colored community attracted to Trinidad by Spain's immigration strategy, by 1789, 2151 whites had come to Trinidad along with 4467 free coloreds (Millette 15).
11. As Kempdoo notes in her email correspondence with the author on August 18, 2004, the character of Bath is based on 'black Muslims' existing in the historical records who 'petitioned [the British government] to be returned to Africa.' In addition, Jean Philippe Baptiste's character is also based 'on a historical character who was a qualified doctor and petitioned the English government to recognize the free coloured rights.'
12. In another story strand dealing with the forced closing of Jean Baptiste Phillippe's medical practice by the colonial administration, he makes this assertion directly while in conversation with his wife, Marie Louise (Elsie's daughter).

References

Dash, J. Michael. 'Introduction' in Edouard Glissant's *Caribbean Discourse: Selected Essays*. Charlottesville: University Press of Virginia, 1989: xi- xlv.

Glissant, Edouard *Caribbean Discourse: Selected Essays*. Charlottesville: University Press of Virginia. 1989.

Hansen, Mark B.N. 'Digitizing the Racialized Body or the Politics of Universal Address.' *SubStance* 104, Vol. 33, No. 2 (2004): p. 107-133.

Millette, James. *Society and Politics in Colonial Trinidad*. Trinidad: Omega Bookshops and London: Zed Books Ltd., 1985.

Munasinghe, Viranjini. 'Redefining the Nation: The East Indian Struggle for Inclusion in Trinidad' *Journal of Asian American Studies* 4.1(February 2001): 1-34.

Murdoch, H. Adlai. '(Re)Figuring Colonialism: Narratological and Ideological Resistance' *Callaloo* 15.1 (1992): 2-11.

Nelson, Alondra. 'Introduction: Future Texts' *Social Text* 71, Vol. 20, No. 2, Summer (2002): 1-15.

Oguibe, Olu. *The Culture Game*. Minneapolis and London: University of Minnesota Press, 2004.

Jewish life in Germany from 1914 to 2004: The story of the Chotzen family

Gorch Pieken

A joint multimedia project of the Federal Centre for Political Education (Bundeszentrale für politische Bildung) and the German Historical Museum documents Jewish life in Germany from the First World War to the present day. Through the personal fate of a German family of Jewish faith, stories of emancipation and integration, persecution and extermination as well as new beginnings and reformation after 1945 and of Jewish life in reunified Germany are told by means of a multimedia application.

In the form of two timelines within a comprehensive data pool, more than 1700 pictures and almost 220 film clips can be viewed. One of the timelines is dedicated to the political and cultural backgrounds to Jewish life in Germany between 1914 and 2004. Running parallel to this is the second timeline which documents the life and fate of a German Jewish family. The Chotzen Family Chronicle begins in 1914 and ends with the death of Eppi Chotzen in April 1992. The political events of these times are given a human face by means of these simultaneous timelines: The political dates and facts of the period are reflected on a personal level in the lives of the Chotzen Family – which comprises of the mother, Elsa and the father Josef as well as their four sons and the later addition of four daughters-in-law.

The Chotzen Family 'Chronical' reveals not just the cruel hopelessness and brutal violence that shatter the daily lives of an otherwise very normal German family, but tells also of the love of a mother for her children. On numerous occasions she begs successfully for her sons and daughters-in-law to be released from the transit camps of Berlin. She uses any means necessary to try and hold the family together during the difficult years of the Nazi dictatorship – in the end for nothing.

The Chotzen Family has a story to tell. It is about companionship and the love of life, as well as the will to survive and the extermination of a German family of Jewish faith.

The Chotzen boys were passionate and proficient photographers, and almost 700 of their photos in eight photo albums have survived. In addition to this, almost 400 cards and letters from Theresienstadt and Riga have survived and are now part of the collection of the Memorial to the House of the Wannsee Conference. A particularly insightful historical document is the large-format housekeeping book kept by Elsa Chotzen between 1937 and 1946, in which every last detail of the daily incomings and outgoings are recorded. Elsa Chotzen's housekeeping book perfectly reflects the family's daily life across nine years in all of its ups and downs, entry for entry.

Computer technology, as opposed to printed text which can only contain letters and words, can be used to add photographs, sounds and films. Alongside letters, drawings, official documents and photographs are the weekly newsreels, private Super 8mm films, excerpts from the evening news and clips from TV and cinema in addition to new film recordings from Theresienstadt and Berlin as well as interviews with survivors and witnesses have been added to the application. Sorted chronologically, the historical events are divided into a multi-layered text system. The layer '0' time strip is always made up of an illustration and a heading. The first layer of information comprises of one or two sentences of explanation. Those who require more information have the option to click into the second layer and, where present, the third and fourth layer where quotations or eye-witness reports can be viewed.

A limited amount of text is permitted on the upper layer, where more users are active; whereas the amount of text present in the lower layers, to which users with more intense interest can channel through to, can be set according to requirements. All texts are spoken by a narrator and are simultaneously subtitled for images and videos. Emphasis has been placed on the user-friendliness and ease of understanding of the navigation: It is straightforward and easy to learn. The icons listed between the forward and back buttons display the number of layers which can be selected. The page that is being viewed is highlighted in red.

Every picture featured in the political as well as the family timeline can be identified by means of a 'roll-over' text which appears.

An extensive glossary and more than 100 individual biographies can be referred to as required from within the application. Each layer of the application has a mini-overview of glossary entries available to the corresponding page to be viewed. This means that abbreviations and specialized terms mentioned in a text can then be viewed in a list in a pop-up menu (which can be reached via the information button) and can then be swiftly interpreted. For example, terms such as 'National Socialism', 'propaganda', 'Gau', 'NSDAP' and 'anti-Semitism'. The same also applies to people mentioned in the text. A simple mouse click causes a short biography to appear on the screen.

Users who aren't familiar with the history of the Chotzen family will find it much easier to get to know the large family. Via the info-button a collage of pictures can be viewed in which Mother and Father Chotzen as well as the four sons and their wives can be seen. Each name is connected to a short biography.

The family chronicle is also illustrated by means of an historical map of Berlin in which all the streets and houses relevant to the family in Berlin are marked with a point. A map of Germany also shows (and links) the places outside of Berlin which were significant to the Family's story. The timeline ends – for the time being – in 2004. The Content Management System makes it possible for further facts and data from future years to be added to the timeline. Even the year 2004 has borne witness to some bad news, but also to some more happy occasions.Progress when using the timelines resembles a probing search or voyage of discovery, or even a reminiscent rediscovery of that which has already been learned. Searching for information becomes a rewarding and advanced journey through time, generating a relationship between the user and the locations of the collection and historical scenes.

For those not wishing to take advantage of the interactive selection there are short films available, each with a maximum duration of three minutes. The subjects which they cover can be selected from the main menu and include the most important aspects of the timelines or with those central points of focus which aren't featured in the timeline, for example subjects such as 'Jewish Life in Germany Before 1914'.

It is of course possible to access all glossary entries or biographies from the main menu. The search function enables fast access to all key terms and phrases. In addition, an alphabetically sorted index also enables a quick overview.

Global information networks and new media and technology have opened up a new quality of learning and new horizons for school children and students throughout the whole world. Special learning modules for school children from primary school and secondary school age can be selected via the main menu.

In the form of subject-oriented guides, channels are opened up through the forest of information contained in the large data pool of the timelines. The apparently limitless capacity to store and link data supports a pedagogically relevant compression of approach to the subject and learning content.Task lists and task sheets to each of the subject guides are available interactively or as PDF-files. In addition to this, school children can still independently research the data pool. With supportive descriptions of the various source genres such as audio, moving image, still image, document, item etc., it is possible for school children to create their own subject-related material and text collections. The PowerPoint presentations planned for lessons can be assembled from own written texts together with objects, documents, images and films from the timeline to create individual essays. To do this, a pupil simply has to click with the mouse on a picture or video and paste the selected image into their essay. The same can be done with text items with which the pupil selects and then combines with their own written text. To view what they have done they can view the collected pictures and text individually at any time.

Pupils can also create their own presentations via a special module available in the menu. This is done by selecting a template. Text can be positioned for example on the right-hand side with pictures on the left or alternatively with pictures at the top and text at

the bottom, etc. Subsequently, images can be selected from the self-created gallery and added or exchanged in the desired order. The same applies to texts copied from the timelines which can then be complimented with reformulated or self-written sentences.

Back in the timeline, the user can then create a comprehensive picture index with details of the title, artist, date and owner/inventory number via the source button.

A free text search enables swift access to all text information and images contained within the application. All items, paintings, graphics and modules available in relation to a keyword like 'Synagogue', for example, can then be viewed.

A search for words and sentences is also possible. By entering a concept like 'school' for example, would lead to all available entries in the timeline being listed. With a single mouse click, the user is brought to the corresponding location in the timeline. A search within the biographies is also of course possible. By entering for example the name 'Adenauer', all entries relating to his life in which he is mentioned by name, as well as his own biography, are displayed.

An important factor is the process for dealing with how the subject matter is structured, the content hierarchy, the presentation of context and the development of a topic focus. This conceptual work demands an intensive comprehension of the subject. The knowledge that each learning success generates goes far beyond the usual receptive forms of learning. Text modules covering the subjects of 'exclusion', 'emigration', 'National Socialism', 'resistance' etcetera are as easily available to primary and secondary school pupils via the main menu as detailed background information to selected items, photos and documents.

Additionally available for teachers are:
– Teachers' guides
– Instructions for use (interactive and as a printable PDF file).
– Answer guide to worksheets etc.
– A list of links and a guide to further literary sources.

For the concept of open participation, users also of course have the option to incorporate their ideas visually or as text.

The full-version of the application 'Jewish Life in Germany from 1914 to 2004' can be ordered in interactive DVD from the selection of publications available from the Federal Centre for Political Education (Bundeszentrale für politische Bildung). Large sections of the multimedia project are available in German and English on the World Wide Web.

The old Montréal heritage inventory database: Toward a renewed collective memory

*Léon Robichaud**

Introduction

The computerisation of heritage inventories has facilitated updates and improved access for heritage information around the world. In the case of Old Montréal, the development of a computerised heritage inventory has also stimulated research and created a core tool around which knowledge about Montréal's historic district can be attached.[1]

Through the use of a web-based database-driven inventory, we were able to resolve several issues which had plagued previous endeavours: updating the information, improving access for a broader range of users, integrating academic research, linking the inventory with other knowledge projects for Old Montréal and bringing our content into the collective memory.

As with any project of this size, our success rests upon the work of several people, bringing together expertise in historical research, heritage management, information management and information layout.[2] In this paper, I will summarize the main components which have made this web site a model in its field.

Objectives

Before the computing era, heritage inventories were often static documents, designed to provide a snapshot of a collection of objects or buildings at a given point in time. Created primarily to assist professionals in the management of the collection and to provide general information for experts in the field, traditional inventories were hard to maintain and offered limited access to the general public.

In 1997, the City of Montréal and the Ministère de la Culture et des Communications du Québec wished to replace an architectural inventory of Old Montréal created in early 1980's with a new database which could be shared by the two instances responsible for the historic district through the Entente sur le développement culturel de Montréal.

The new database was to be a knowledge repository and a base from which the information could be disseminated to the public. The main users would still be professionals from governmental agencies with occasional users from universities and other researchers. Eventually, this database would be made available to the public. Given its broad objectives and the technical obstacles to sharing data between two very different systems, it became clear that the best solution to these general objectives was to move right away to a database driven web site. As a result, the public gained access to the data sooner than originally planned.

As far as information was concerned, different types of information had to be included in the inventory: buildings, public works of art, biographies, events, streets and squares, as well as a bibliography of documents on the historic district. Although the buildings were to be a major focus of research, the overall inventory had to reflect the many aspects of the history of Old Montréal.

Given the project's general objectives, the Old Montréal heritage inventory was not to be a hierarchical system centered on the building. While the buildings remain at the core of the inventory, the system had to

* Université de Sherbrooke

1The Old Montréal web site can be found at http://vieux.montreal.qc.ca, the inventory itself is at http://vieux.montreal.qc.ca/inventaire/hall.htm .

2This project was initiated in 1997 by Gilles Lauzon, with the Société de développement de Montréal, and Madeleine Forget, of the Ministère de la Culture et des Communications du Québec. The system was developped by Léon Robichaud and Alan M. Stewart of Remparts-RS. Many others have contributed to the project, for the complete list, see http://vieux.montreal.qc.ca/credits.htm

offer a broader approach with multiple points of access. We had to be able to query the database from any of its components (buildings, public works of art, biographies, events, streets and squares, documents). Futhermore, related components had to be linked together to provide transparent navigation to the user. The structure had to be flexible enough to incorporate later additions. Finally, apart from a variable-based search engine, access points would be created to provide a semi-structured form of navigation based upon space (sectors of the district) and time (major historical periods and key years).

Content and structure

The database offers information on 584 existing buildings.[3] We also offer information on 230 individuals and 132 institutions or corporations. The database contains 441 events (brief notice), records for 17 public works of art and information on 78 streets, squares and wharfs. The extensive bibliography contains over 1770 records on publications and archival collections related to Old Montréal.

The data is stored in a MySQL database, using a very broad relational structure. There are 7 main components to the structure: buildings, people, groups, events, works of art, public spaces and bibliography. Each of the components has a main table and a number of dependent tables in a one-to-many relationships. Different elements from different components (buildings and people) as well as different elements from a given component (people and people) are linked together via linker entities in many-to-many relationships. As a result, the database, including reference tables, includes over 80 individual tables.

Navigation

When a user enters the heritage inventory,[4] several modes of access are offered, with spatial navigation being at the forefront. A flash version and a text with images version of this portal are available, depending on the browser configuration.

With spatial navigation, users can click on a sector of the district to an image-map of the sector (with hyperlinks to buildings, streets and works of art) as well as a list of all buildings, people, groups, and works of art associated to the sector, grouped by period. A click on the the hyperlink for an individual element leads the user to an individual record.

Two modes of temporal navigation are available, by period and by key year. Using the same underlying data, the page for a period will offer a list of events, people and groups, as well as buildings, streets, works of art and documents associated with the period. For buildings, we indicate those which were erected during the period and those which were significantly transformed during the period.[5] When users click on a key year, information is focused upon the key events, people and buildings associated with a year which was significant in the history of Old Montréal. For example, 1785 provides a snap-shot of the situation following the American War of Independence, during which Montréal was occupied by rebel troops. During the year 1849, the Parliament buildings were destroyed by fire while 1873 is at the heart of Montréal's industrialisation and the beginning year of an economic depression.

For users who have more specific requirements, several modes of querying are available. Buildings can be searched by name, by street address, by historical and functionnal criteria and by architectural criteria. People and groups can be searched by period and by role in the urban development of the district. Other tools are available to search events, public works of art, streets and squares and documents. Query results are presented as a list of hyperlinks to individual records.

Building records offer current and historical images. In a few cases, elevations are also available. The following section presents the building name(s), the location (address and plan), the main physical features (with a link to a detailed list), and, if applicable, association to a building grouping. For most buildings, there is a text history, and for all buildings, the construction and main transformations are indicated with

3 The total number of buildings in the district has remained close to 585. Since the creation of the database in 1997, 11 buildings have been destroyed by fire or were demolished for safety reasons. There have been nine new buildings, some of which have replaced those destroyed by fire. The information on the 11 lost buildings remains in the system but is not available to the public.

4 http://vieux.montreal.qc.ca/inventaire/hall.htm

5 Our definition of a building, of the date of construction and of a significant transformation, as well as several other expressions, are presented in the glossary.
http://vieux.montreal.qc.ca/inventaire/doc/aide/glo.htm

the building's function. There are also references to architects, builders, owners and occupants with hyperlinks to major figures. Heritage protections are also listed, indicating the level of government responsible for the type of protection.

Individual biographical records indicate dates of birth and death, provide an image when available and offer a narrative of the individual (or couple) anchored in a key year. As a result, there can be more than one narrative for a given individual, providing windows into that person's involvement with Old Montréal at different points in time. There are also links to other individuals, to corporations, to buildings and to sectors, always with bibliographic references.

Institutional and corporate records are also anchored to key years. They provide the dates of formation (and dissolution if applicable), a list of major members for the key year and a narrative. There are also links to individuals, to other corporations, to buildings and to sectors, with bibliographic references. When available, illustrations are provided.

Records for public works of art offer an image of the monument in its environment as well as a detailed view. The record indicates the title of the work, author(s) and collaborator(s), date and location for the installation, as well as the current owner. The narrative provides a history, a description and an interpretation of the work of art. Links to authors or to persons represented are provided.

For public spaces (streets, squares and wharfs), different views are presented from different periods with a narrative of the street's history and links to the different sectors which the street crosses in Old Montréal.

Scripting, editing and usability

The database content is managed via online input forms, written in perl. The user name and password defines the level of access and the components which are available for editing. Content is validated on the form with javascript and before updates with perl scripting. All editing is performed online by our team of consultants spread across the Montréal region.

Links between components are added manually by the consultants because links between a building and an individual are not always obvious and cannot be automated. All links are bidirectional and the consultant therefore does not have to repeat the linking process from the receiving end.

When a large number of images need to be added, they are uploaded via FTP and the related information is added to the database via a batch script. Individual images are integrated via online forms.

Our «internal» community of users has very different levels of expertise with regards to computing and the web. As a result, we obtained feedback very quickly on the site's usability. A few features were added in response to this feedback, namely the spatial mode of navigation. A user-survey form was added to the site in 2004, which was surprisingly successful in providing us with more feedback on the site and on its usability from the general public. We work on maintaining ease of use while adding more powerful functionalities and are constantly updating its look.

Our public interface was recently overhauled with both a new «2005» look and a switch from perl to php. We are also moving towards XHTML, but our pages do not yet comply with W3C standards.

Results

Compared to approximately 100 experts who consulted the paper-based inventory every year, the web-based inventory reaches over 10,000 unique users every month. Our broad user community includes heritage professionals, architects, developers, building owners and tenants, real estate agents, historians, students, genealogists, and people generally interested in Montréal's historic district.

On the academic side, new information resulting from the site has stimulated research. A collective work was published in 2004 on the history of the heritage district focusing on its surviving heritage.[6] Another research project was launched in collaboration with the Université du Québec à Montréal to provide more insight on a particular building type from the late nineteenth century: the *magasin-entrepôt*. This mixed-use building typically had a showroom on the first floor with warehousing and even light manufacturing on the upper floors and was the dominant form of new architecture in the district for a 30 year period. As new knowledge raises new issues, this site can serve as a base to explore other aspects of Old Montréal's history in the future. We are also currently evaluating the

6 Gilles Lauzon and Madeleine Forget, eds, *Old Montréal: History through Heritage* (Québec, Les Publications du Québec, 2004).

possibility of integrating 3D models of certain buildings into the web site.

Conclusion

We needed a powerful tool to manage a diverse content while providing easy to use tools for the public. The Old Montréal heritage inventories have successfully reached its original traget audience and has broadened its appeal to an even broader public. The site has become the reference for online information on Montréal's historic district and has stimulated new research. A model for other projects, our database has also become the standard platform upon which the city's heritage section provides information for all of Montréal's heritage sites.[7] Since 1998, when the site was launched, a solid ressource is available to disseminate information about Old Montréal for the media, for guides, and for users in general.

7See http://patrimoine.ville.montreal.qc.ca/inventaire/

On the ground and '6 feet under'

Hartmut Tschauner & Viviana Siveroni Salinas***

Mobile GIS and photogrammetric approaches to building 3D archaeological spatial databases in the field

Introduction

GIS technologies have been available for decades and, while still far from universally adopted in archaeology, have been in use by individual practitioners for a long time. Archaeology's traditional emphasis on graphical documentation has always implicitly recognized the primacy of the spatial dimension of the archaeological record: time is derived from space (stratigraphy) and our understanding of past cultural systems is based on the spatial relations between features and artifacts (context). That the potential of GIS in such a heavily 'spatialized' discipline has not been fully realized is due to archaeologists' failure to recognize GIS as a set of general-purpose spatial data management tools that needed to be adapted to the core needs of their field. Instead, archaeology has tended to import, along with the spatial data management tools, specific applications of those tools from other fields that were earlier adopters of GIS. In fairness, larger and better funded fields represent sufficiently lucrative market segments to sustain customized commercial software products. Moreover, in-house mathematical and computer science expertise permits the development of highly specialized software tools. As a small, notoriously underfunded, and largely non-quantitative field, archaeology is at a distinct disadvantage relative to those fields.

GIS and archaeological field data collection: more than a matter of convenience

One core area of archaeology – arguably its defining feature in the public eye – where custom-tailored spatial technologies could be extremely productive and address pressing issues of the discipline is primary data collection in the field. Going back to the (still paper-and-pencil-based) spatial archaeologies of the 1970s, most archaeological uses of GIS have skipped to 'sexier' analytical applications. The availability of data appropriate for such analyses has been taken for granted. Moreover, using generic tools and borrowing models from other fields, the overwhelming majority of applications are two-dimensional, and regional analyses far outnumber intra-site studies.

However, attempts to perform GIS analyses on spatial data from archaeological publications will frequently run into serious problems with data quality. The ability of GIS to integrate layers of information from many sources brings these previously ignored accuracy issues to the fore. Furthermore, collection and documentation of archaeological spatial data is notoriously selective so that the evidence is often insufficient to support formal analyses of the spatial dimension of the archaeological record. In particular, profile drawings of stratigraphic sections at the edges of excava-

*Department of Archaeology, Seoul National University, Gwanak-gu, Sillim 9-dong San 56-1, Seoul, 151-742, Korea
**Department of Anthropology, 3126 Wesley W. Posvar Hall, University of Pittsburgh, 230 South Bouquet St., Pittsburgh, PA 15260, U.S.A.

tion units offer a spotty and arbitrary sample of a site's stratification. Save a few spot elevations on top of 'significant' features, between sections the vertical dimension goes essentially undocumented. Needless to say, the substantial labor involved in digitizing paper records could be saved if digital data were collected in the first place. Thus, although probably no other activity consumes more field time than the graphical documentation of spatial information, that information is notoriously incomplete, can be of dubious quality, and is buried in a paper medium that deprives it of the essential quality of 'data,' that is, being susceptible to analysis linked to a particular theoretical question or framework. These spatial pseudo-data, along with numerous domains of descriptive (attribute) and image information, are collected separately in multiple places and formats – a hodgepodge of lists, forms, descriptive texts, drawings, and photographs, which hinders data integration and holistic cross-examination of all classes of evidence.

To make matters worse, an increasingly small fraction of this limited-use, view-only documentation of archaeological spatial data collected at the source makes it onto archaeologists' desks. Costly graphic documentation and lengthy tabulations or descriptions are usually the first parts of a manuscript to be sacrificed to economic considerations, and this publishing bias sends a clear message that cannot fail to have an impact on field data collection: the academy rewards creative interpretation rather than 'mere' collection of evidence, your conclusions are more important than the quality of the evidence that supports them. In a discipline that inevitably destroys its evidence in the process of studying it, this creates a real and present danger of archaeology devolving into a pre-scientific state of knowledge claims largely being justified by an author's personal credentials, as well as the arguments' logical consistency, and hardly being backed by solid evidence *that others can challenge* (cf. Barker 1993:13-14) Before this backdrop, the universal use of digital spatial technologies in the field, both for more complete collection of spatial data and as a prerequisite for their electronic publication in full and in a readily analyzable format, takes on enormous urgency, not as a mere technical improvement, but as a matter of significant theoretical repercussions.

Units of archaeological stratification are a class of spatial data sufficiently different from those of other disciplines to require specialized modeling approaches that are not part of the standard GIS repertoire and need to be adapted from existing methodologies, mostly from geology, or developed from scratch. Like geological strata, archaeological units are volume solids and ought to be modeled as such. Unlike geological strata, archaeological units are typically not interpolated between sparse sampling locations (boreholes), but fully exposed by horizontal excavation. Current GIS data structures do not model solid volume entities; they are at best 2.5 dimensional, with an elevation attribute tagged onto two-dimensional features and all spatial analysis taking place in a horizontal plane. To these models, the superimposed strata in an excavation unit are spatially identical, as long as they cover the same horizontal extent. The few geological packages that have pioneered a voxel (*vol*ume pi*xel*) approach to solid modeling are geared towards 3D interpolation and have trouble creating accurate stratigraphic units from fully exposed and mapped interfaces. Published applications of GIS to intra-site analysis and archaeological excavation reflect this software situation, dealing mostly with architecture or horizontal distributions of artifacts susceptible to spatial analysis in two dimensions (*e.g.*, Buck*, et al.* 2003; Craig 2000; Green*, et al.* 2002; Levy*, et al.* 2002; Peretto*, et al.* 2001), while vertical, stratigraphic, and truly 3D analyses are rare (*e.g.*, Nigro*, et al.* 2003; Spikins*, et al.* 2002).

From the ground into the spatial database: mobile GIS at Huayurí

This section briefly sketches our approach to creating complete, electronically publishable, and analyzable, 3D models of excavated sites. This is not meant to be a guide to cutting-edge digital field methods; on the contrary, it is a practical description of simple, improvised procedures that we developed on a shoestring budget during excavations at the site of Huayurí (Ojeda 1981), a 20-ha agglutinated habitation site in the Santa Cruz Valley on the south coast of Peru dating to the Late Intermediate Period and Late Horizon (A.D. 1000–1532). Since archaeologists seldom benefit from the cost-saving aspects of digital technologies – given that so much of the labor used in the field is free or severely underpaid – even the improvised solutions described here come at a substantial cost.

Stratigraphic recording: contact topography

Citing the destructive nature of excavation, archaeological textbooks have always admonished excavators

to record the full 3D structure of their sites, although until recently this could only be reported through multiple, flat drawings and descriptive text. In reality, however, stratigraphic recording through section drawings at ultimately arbitrary locations (edges of excavation units, balks) and spot elevations is highly selective and simply does not produce the data required for full 3D models. Now that technologies exist to build and distribute the models and to analyze them beyond purely visual inspection, our recording strategies need to adjust to make sure we collect the necessary data.

Although we ultimately want to build solid volume models, the most economical way to do so, both from the fieldwork and software perspectives, is surface-based contact topography, i.e., microtopographic maps of the upper interfaces of all units of stratification. Standard 2.5D GIS software is able to create and display the interfaces; specialty software subsequently creates solid volumes between the interfaces. Cut features (pits, postholes, etc.) are treated like layer interfaces. Their top interface is outlined and mapped on the surface from which they were cut; their bottom interface is mapped after they have been excavated.

A total station is used to map simple interfaces. The point density depends on the nature of the interface. An irregular occupation surface may require a very high density of points to model all its details and be able to relate them to artifact displacements and other formation processes. As a spatial data collection task, stratigraphic recording has extremely high accuracy requirements, particularly in the most difficult vertical dimension. Layers of less than 1 cm thick may represent meaningful stratigraphic events and therefore will have to be distinguished. Complete systematic interface mapping produces a substantial volume of point data, typically in the tens of thousands each season, and hundreds or thousands of 3D polygon features.

More complex interfaces are recorded by photogrammetric means. 2.5D surface models are extracted from sets of oblique digital photographs taken from several angles and ground-referenced with a few total-station points. Photogrammetry programs designed for industrial product development, accident-scene reconstruction, and architectural applications are relatively inexpensive and appropriate for the scale of most excavation units (cf. Green*, et al.* 2002). Accurate, detailed surface models require a dense grid of standardized, high-contrast targets in the images. We use circular, reflective targets mounted on chips of heavy plastic material. We produce these chips ourselves from commercially supplied rolls of adhesive target tape. A more elegant and wind-proof alternative is a target projector, but these devices are costly and do not operate on battery power. An even costlier option is 3D laser scanning, but it remains to be seen how the scanners cope with the diffuse reflectance of earth surfaces and point cloud software with highly irregular shapes that cannot be constructed from geometric primitives. The latter problem also arises when building surface models from hundreds of photogrammetry or smaller numbers of total-station points. Most surfacing algorithms will require some manual operator input.

Recording features and occupation floors

As described above, cut features are generally treated like any other unit of stratification; they are outlined in plan view on the surface from which they were cut, and their bottom interface is mapped once they have been excavated. Some cut features may pose special problems, however, in particular small ones (such as postholes) and those with overhangs (*e.g.*, storage pits), in which it may be impossible to position reflector equipment or photogrammetry targets. At complex sites with multiple occupation phases, such as Huayurí, the sheer number of features may also represent a considerable challenge to complete 3D recording.

Stepwise photogrammetric documentation provides a solution to these problems, simultaneously producing 3D models of each interface as well as slices – both vector models and orthophotos – through any features cut from or into it. 3D models of the cut features are then built, like in computer tomography, from multiple slices or outlines on successive stratigraphic interfaces that the features intersect.

3D modeling and spatial analysis: voxels and solids

Current GIS software models interfaces as stacked surfaces, as they are mapped in the field. The purpose of stacked surfaces is mostly on-screen visualization. While the communicative power of such models is not to be belittled (rotating in 3D, peeling off layers, etc.), their analytical potential is limited. They serve as the basis for volume calculations and automatic section generation, usually restricted to one surface at a time. They do not support 3D spatial analysis of the volume solids sandwiched between the surfaces.

For example, we might want to find volumes that have a specified density of pottery remains and level of magnetic anomalies and how their distribution relates to occupation surfaces or features. Or we might investigate how some geochemical concentration measured at sparse sampling locations is related to the site's stratification by interpolating the geochemical concentrations in 3D space and intersecting the resulting 3D distribution with the interface-mapped volume model of the site. Or is there a significant association between the densities of two artifact categories in 3D space? How do artifact densities or other measurements behave in relation to 3D distance from certain features? Similar analyses are the bread-and-butter issues of two-dimensional GIS. For archaeological applications, particularly formation-process research, they need to be extended to 3D solids.

At present, there are two alternative data structures for 3D solid modeling: CAD 3D solids and voxels (*volume pixels*). The difference between the two is analogous to that between vector and raster GIS in 2D.

CAD solids are the main data structure of industrial design applications. Therefore, the editing techniques supported by these applications are tailored to symmetrical shapes. Complex models are generated by rotating 2D vectors or Boolean operations on 3D primitives. However, the creation of arbitrary, non-symmetrical solids modeling archaeological strata is easily automated by extruding grid cells modeling the lower interface into vertical, prismoidal columns touching the upper interface (or vice versa) and merging the columns into a single complex solid. Like its 2D analog, vector GIS, this approach supports feature-based spatial overlays through Boolean operations (intersect, subtract, union) and attaching unlimited attribute data by linking external database records to the solid features. However, Boolean operations are extremely computationally intensive and in a CAD environment do not extend to the attached attribute sets.

Voxel models are the 3D analogs of rasters, 3D matrices of extruded, cubic pixels. This data structure was developed for game design and medical imaging. Like rasters, voxels are appropriate for modeling continuous distributions of variables in space, and this is how they are used by some geological software packages, which perform 3D interpolation of observations and measurements made in boreholes. These algorithms do not create realistic stratigraphic models from data points collected on upper and lower interfaces. As a workaround, undifferentiated solid models of an entire excavation block may be filtered between bounding grids, i.e., voxel values between interfaces may be replaced with a stratum number. Since this slicing procedure needs to be repeated one unit at a time, building a complex stratigraphic column is a tedious process. Spatial analysis of voxel models is accomplished through voxel-by-voxel arithmetic operations (model vs. model or model vs. scalar) analogous to 2D map algebra.

Outlook: software developments

While existing GIS software offers no support for 3D spatial analysis, there are specialized civil engineering and geology packages that do. However, these applications have trouble dealing with the complexity of archaeological stratification, they tend to be costly, and mainstream archaeological users may not be willing to learn a whole suite of specialized software applications. Therefore, HT is working on two software packages that will address specifically archaeological requirements of 3D spatial analysis and will be made available to the archaeological community at nominal or no cost.

The first is a multi-surface section and fence-diagram tool for the ArcGIS environment. Civil engineering applications include tools that can generate sections through multiple, superimposed surfaces, but they tend to break down when confronted with the numbers of strata commonly encountered at archaeological sites. At Huayurí, for example, an excavation area of roughly 250 m² so far has produced 358 units, and this number will likely have doubled when the analysis is completed. The tool under development will create sections and fence diagrams of arbitrary complexity as vertically oriented, to-scale vector polygons at user-specified locations in the site model. It will also have the ability to orient the model so that a flat view of a section may be printed or exported as an image.

The second, more ambitious project is a voxel modeling system for archaeological stratigraphy that will automate the creation of complex voxel models for 3D spatial analysis from stratigraphic interfaces mapped by the contact topography method and stored as vector entities in CAD formats (as commonly produced by surveying software running on mobile devices), GIS vector features, and raster surfaces. 3D map algebra is massively simpler to program than topological opera-

tions on 3D solids and is appropriate for both discrete features and continuously distributed variables. The major challenge will be optimizing the use of machine resource since complex 3D modeling requires enormous amounts of CPU time and memory.

References

Barker, P. 1993 *Techniques of archaeological excavation*. 3rd ed. Batsford, London.

Buck, P. E., D. E. Sabol, and A. R. Gillespie, 2003 Sub-pixel artifact detection using remote sensing. *Journal of Archaeological Science* 30(8):973-989.

Craig, N., 2000 Real-Time GIS construction and digital data recording of the Jiskairumoko excavation, Peru. *SAA Archaeological Record* 18(1):24-28.

Green, J., S. Matthews, and T. Turanli, 2002 Underwater archaeological surveying using PhotoModeler, Virtual Mapper: different applications for different problems. *International Journal of Nautical Archaeology* 31(2):283-292.

Levy, T., R. Adams, A. Hauptmann, *et al.*, 2002 Early Bronze Age metallurgy: a newly discovered copper manufactory in southern Jordan. *Antiquity* 76(292):425-437.

Nigro, J. D., P. S. Ungar, D. J. d. Ruiter, *et al.*, 2003 Developing a geographic information system (GIS) for mapping and analysing fossil deposits at Swartkrans, Gauteng Province, South Africa. *Journal of Archaeological Science* 30(3):314-324.

Ojeda, B., 1981 La Ciudad Perdida de Huayurí. *Boletín de Lima* 16-18:78-82.

Peretto, C., M. Arzarello, F. Fontana, *et al.*, 2001 Excavation in progress: the Palaeolithic site of Isernia La Pineta, Italy. *Prehistoria* 1:138-149.

Spikins, P., C. Conneller, H. Ayestaran, *et al.*, 2002 GIS based interpolation applied to distinguishing occupation phases of early prehistoric sites. *Journal of Archaeological Science* 29(11):1235-1245.

World museums on the internet: A brief overview

*Timur Valetov**

The Internet could be considered a virtual library giving easy access to almost any kind of information. However, the quality of this information depends on the effort that were made by the teams creating the websites. Some organizations put a lot of information on the web including electronic texts, databases, and image collections. Some only put their contact information. The value of the information depends on the subject, so it is necessary to improve especially those of the organizations which are willing to share information. For historians, these are world heritage institutions such as archives, museums, and libraries. This paper gives a brief overview of world museums on the Internet. We shall try to describe different types of museum sites, and compare Internet branches of museums in different countries.

We prepared this study by compiling a museums links list. The links list is now available on the web site of the Faculty of History at Moscow State University.[1] It could be easier to use any other links list, however we found many problems with existing examples. It is necessary to say that one of the best museum links list is the list held by ICOM (the International Council of Museums).[2] Their huge and useful links collection is the basis for some other links lists, but it seems to have some serious problems.

Firstly, it was made several years ago, and it contains a lot of outdated information. It is a kind of problem common in a lot of links collections, and the problem is serious, because new museum sites appear rapidly. Secondly, ICOM's collection includes both official websites and non-official web-pages placed elsewhere, and these non-official pages can contain just brief information with a few images of a low quality. However, a non-official web site can also be impressive, and it sometimes happens that one museum has many websites, each of which is unlikely to be official in any way.[3] Finally, ICOM's collection doesn't sort the museums by importance as centers of world culture but sorts them alphabetically. This is why famous museums, such as the British Museum or the Louvre, can be placed somewhere in the middle of a long list after some virtual galleries of practically unknown artists. For example, we can check the list of Russian museums, and the most important national museums are placed (among 515 links) under the following numbers; 373 the State Hermitage; 378 the State History Museum; 390 the State Russian (art) Museum. The Tretyakov Gallery is represented by five links, among them the official web site is link #394, and at the same time two other national museums, one of which is the National Museum of Oriental Art, are not represented at all. In contrast, there are many less valuable websites higher up in the list; 1 the Transsib Railroad historical photo gallery (not responding); 10 the virtual exibition (20 photos without any text); and 13 is a text page telling about the invention of radio. Of course, it is very difficult to create and to maintain a good collection of links because it is necessary to understand the importance of different museums in each country of the world. Nonetheless, it should be done, because too many links hide the most important museums in ICOM's (otherwise useful) list.

We can find the same problem reflected in another huge collection of museum web site links available on the web site of the Tamkang University in Taipei.[4] The list is very long and it doesn't matter whether the museum has a large site or only a single web page.

*Moscow State University

1 http://www.hist.msu.ru/ER/museum_e.htm

2 Main address is http://icom.museum/, links pages – http://icom.museum/vlmp/world.html.

3 For example, the National Archaeological Museum (Naples) has at least two well-presented sites: http://www.archeona.arti.beniculturali.it/sanc_en/mann/home.html and http://www.marketplace.it/museo.nazionale/ . Besides, there is also the third webpage of this museum – http://www.cib.na.cnr.it/mann/ind.html.

4 http://www.lib.tku.edu.tw/museum/engmuseum/engnetsource.htm

Museum sites in different countries

As the Internet itself, the development of a museum site differs very much from one region of the world to another. The best situation is probably in the USA where most local museums usually have a very good Internet site. The worst situation seems to be in Africa (except Egypt and South Africa), which is understandable, considering the finance and general development of the museums themselves. As an illustration of the development of museum websites it is possible to count official wesites of museums in different regions. In our link collection we found:

Western Europe: 146 sites;
Eastern Europe: 58 sites;
Russia: 71 sites (more than in any other region were counted because we included all the museum websites);
The US and Canada: 35 sites (plenty of museums were not included);
The rest of the Americas: 73 sites;
Western and Central Asia: 25 sites;
China, Japan, and Korea: 33 sites;
South and South-Eastern Asia: 19 sites;
Australia and the Pacific: 20 sites;
Africa: 20 sites.

It is often possible to find web-portals of national councils of museums. In the countries where museum websites are very good themselves, the national councils are the coordinators of the Internet activities and they usually support very good lists of museum links. In some countries these central museum sites contain the information related to all the national museums – good examples include the Ministry of Culture in Greece, the National Board for Cultural Heritage in Singapore, and the Mexican National Institute of Anthropology and History.

The principal language of world museum websites is, inevitably, English. This is why we seldom find a museum web site which has a second-language version (except the native language) that is not an English version. However there are a lot of sites which have the native language version only, and inevitably it greatly reduces the importance of these sites, especially in case of 'exotic' languages such as Japanese or Russian. Almost all the national museums (except perhaps a few located in Latin America) have an English version.

To summarize we can say that museum sites are very well developed in the USA and in some European countries (like the UK, the Netherlands, Switzerland or France). Museums in Japan, the Republic of Korea, Taiwan, Singapore, Canada, South Africa, Australia, New Zealand and almost all Europe are also well represented on the Internet. We also see rather good representation in many countries in Central and South America (although most of them are in Spanish only), China, Iran, the Arab Emirates, Israel, and Egypt. India, some countries of South-Eastern and Central Asia, as well as countries of the Caucasus produce some museum websites, but not enough to make significant impact. Museums of some countries (among them are Turkey, Albania, Pakistan, Tajikistan, North Korea, Laos, Cambodia, Malaysia, and many countries of Africa and the Caribbean) are not represented at all.

Structure and design of museum websites

Museum websites are all of a similar kind with similar information, so they usually have a standard structure. The nucleus of a site is information concerning the museum's permanent collection and temporary exhibitions as well as visitor information and the history of the museum. It is often possible to find a brief overview of the museum's permanent collections, but we rarely find information about possibilities to research the museum's collections or to study in the museum's library. It is often possible to find the museum hall plan; we can usually also find some images of objects, but their quality and quantity vary. Such information as museum research publications, or collection databases are rarely presented.

Museums attract most of their visitors by their rich collections with many famous masterpieces, ancient rarities and precious jewels. This is why museums have an additional advantage in making the websites very attractive which libraries and archives usually do not have. It is an opportunity to make a virtual gallery of images, and some museum sites contain a lot of images. They sometimes use a lot of multimedia in their design, for example, on the Hoam Museum web site (near Seoul).[5] However, a lot of multimedia animations are good just for a short virtual visit and can annoy those who try to work with the site information. Museum websites often use virtual panoramas, which are essential for cultural sites or for images of

5 http://www.hoammuseum.org

Table 1. Summary of the overview of museum websites from different world regions

	How many images are presented in the virtual collection			
	No virtual collection	Some dozens	A few hundreds	More than 500
Western Europe	1	0	2	7
The USA and Canada	1	1	1	7
Eastern Europe	7	2	1	0
Russia	2	3	2	3
Asia	2	5	1	2
Australia and Oceania	7	3	0	0
South America	3	7	0	0
Africa	6	3	1	0

the museum halls. Perhaps the best example of such panoramas is on the 'World Heritage Tour' project of UNESCO[6].

Virtual collections differ very much from one museum to another. We made a short overview of the digital collections of some museums from different world regions and have presented it in the appendix tables. It shows that only some museums from Western Europe and Northern America present many images (see table 1).

Let's analyse selected examples. The famous Egyptian museum in Cairo[7] presents only 189 images in its digital collection, and the quality of images is very low – about 100x200 pixels (see figure 1). But the images are accompanied by short research articles related to the objects. The Tokyo National Museum, which holds only 136 objects in its digital collection, has some objects presented by images made from different angles and it is possible to enlarge images to about 800x1200 pixels (see figure 2).

The research activities of museums are seldom presented on the Internet. Websites are mostly briefly illustrated overviews of the museums, and research papers or collection databases are very rare.

However, we can find some museum websites which are very good examples of combining beautiful images and research articles. For example, the Metropolitan Art Museum (New York) contains a large database of images (about 5 thousand) that are combined into thematic collections. They also created a brilliant course of lectures on art history in the project 'Timeline of Art History'. It covers all of world art history and all the topics are illustrated by maps, chronological scales, and of course by the images from their own database. Another very good example of a large image database accompanied with research articles is the 'Compass' system by the British Museum.

We can find also a few national projects which try to collect the general database of the national museum collections, like the 'Joconde' project made by the Ministry of culture in France.[8]

To conclude, we can say that virtual travel through the museum websites of the world is very interesting – it allows us to observe a lot of world cultures, and the Internet gives a very nice opportunity for this. Each museum web site is interesting, but it is much more pleasant and useful to visit large sites with good content. However, only a few museums present their research activities online, so the construction of virtual collections is far from finished. Maybe it will lead to a large database of international world heritage collections. As the Internet grows rapidly, so hopefully we shall see the progress.

6 http://www.world-heritage-tour.org. Requires QuickTime.
7 http://www.emuseum.gov.*e.g.*
8 http://www.culture.gouv.fr/documentation/joconde/fr/pres.htm

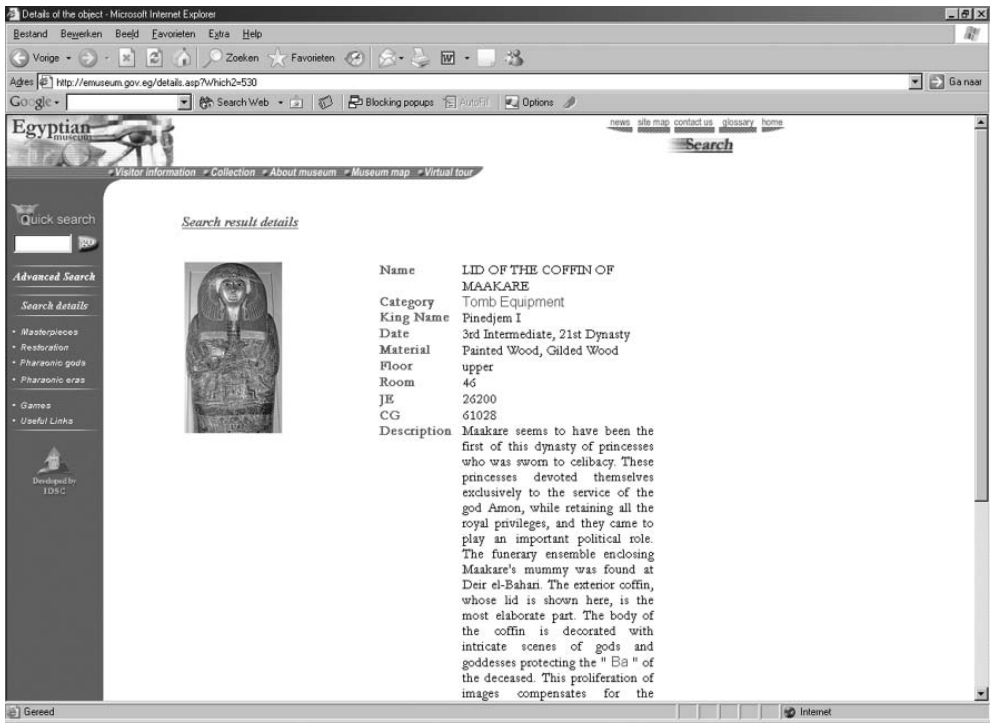


Figure 1. An example of a record in the digital collection. The Egyptian Museum, Cairo.

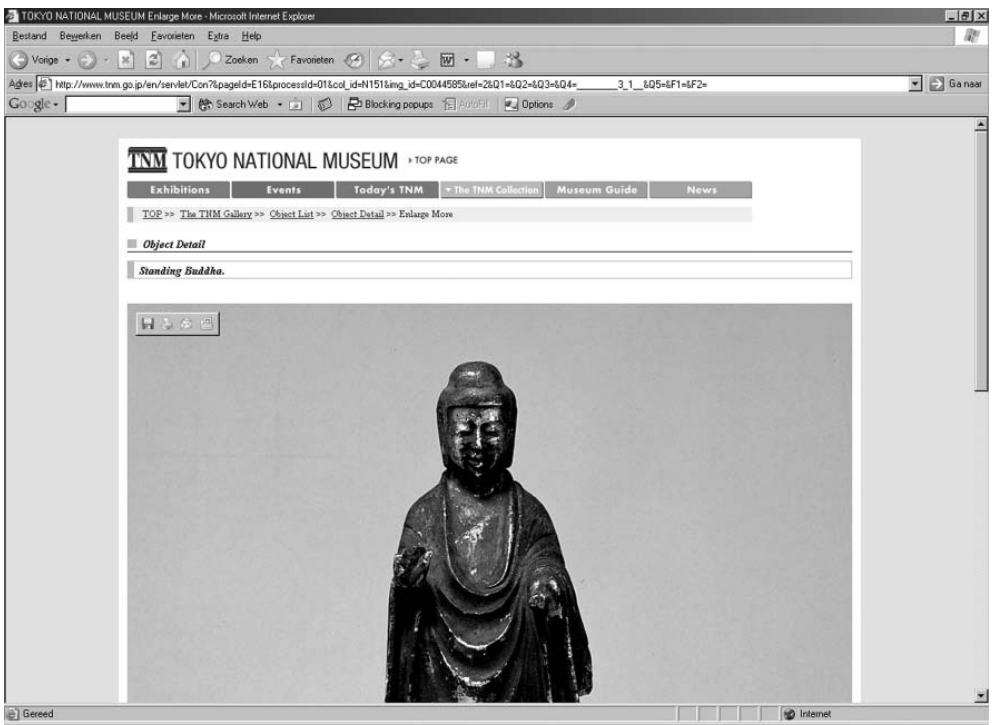


Figure 2. An example of a record in the digital collection. The Tokyo National Museum.

Appendix. The brief overview of some selected museums from different world regions (dark-grey background marks the museums which were not counted for the table 1.)

		Languages (except native and English)	Is there Search on the web site	Are there articles about thematic collections and halls	Is there a digital collection	How many images are presented	Are there articles related to the selected objects
Western Europe							
The British museum	www.thebritishmuseum.ac.uk	-	yes	very brief	yes	5500	yes
The National Museum of Antiquities, Leiden	www.rmo.nl	-	no	yes	no	25	yes
The National Museum of Ethnology, Leiden	www.rmv.nl	-	yes	yes	yes	1000	no
The Art History Museum, Vienna	www.khm.at		no	no	yes	approx. 600	yes
The National Gallery, London	www.nationalgallery.org.uk	-	yes	yes	yes	approx. 10000	yes
The Tate Gallery, London & Liverpool	www.tate.org.uk	-	yes	yes	yes	hundreds	yes
The State Museums, Berlin	www.smb.spk-berlin.de		no	yes	no	no	no
The Bavarian National Museum, Munich	www.bayerisches-nationalmuseum.de		no	yes	no	no	no
The National Museum of Ireland, Dublin	www.museum.ie	-	no	yes	no	no	no
The Prado, Madrid	museoprado.mcu.es		no	no	yes	approx. 100	yes
The Vatican Museums	mv.vatican.va	Germ, Span, French, Port, It	yes	yes	no	no	no
The Uffizi, Florence	www.uffizi.firenze.it		no	no	no	no	no
The Louvre, Paris	www.louvre.fr	Span, Japan	yes	yes	yes	approx. 600	yes
The Musee d'Orsay, Paris	www.musee-orsay.fr	Spanish	no	yes	yes	approx. 500	yes
The Musee Guimet (Oriental Art), Paris	www.museeguimet.fr	Chinese	no	yes	yes	approx. 100	yes
The US and Canada							
The Metropolitan Museum of Art, New York	www.metmuseum.org	-	yes	yes	yes	5000	yes
The National Gallery of Art, Washington	www.nga.gov	-	yes	yes	yes	more than 5600	yes
The Museum of Fine Arts, Boston	www.mfa.org	-	yes	yes	yes	approx. 13600	yes
The Fine Arts Museums of San Francisco	www.thinker.org	-	yes	yes	yes	thousands	yes
The Museum of Modern Art (MoMA), New York	moma.org	-	yes	yes	yes	approx. 200	yes
The Guggenheim Museum, New York	www.guggenheim.org	-	yes	no	yes	approx. 500	yes
The Art Institute of Chicago Museum	www.artic.edu/aic/	-	yes	yes	yes	approx. 500	yes
The Detroit Institute of Arts	www.dia.org	-	yes	yes	yes	more than 1100	yes
The National Gallery of Canada, Ottawa	national.gallery.ca	-	yes	yes	no	less than 100	no
The Royal Ontario Museum, Toronto	www.rom.on.ca	-	yes	yes	no	no	no

Eastern Europe							
The National Museum of History, Sofia	www.historymuseum.org	-	yes	yes	no	no	no
The National Museum, Valjevo	www.museum.org.yu	-	no	yes	no	no	no
The National Museum of Slovenia, Ljubljana	www.narmuz-lj.si	-	no	yes	no	no	no
The Museum of Archaeology, Zagreb	www.amz.hr	-	no	yes	yes	less than 50	no
The Museum of Ethnography, Zagreb	www.etnografski-muzej.hr	-	no	yes	no	no	no
The Hungarian National Museum, Budapest	www.hnm.hu	-	no	no	yes	more than 60	yes
The National Museum, Warsaw	www.mnw.art.pl	-	no	yes	no	no	no
The National History Museum of Transilvania, Cluj-Napoca	www.museum.utcluj.ro	-	no	yes	no	no	no
The National Museum, Praha	www.nm.cz	-	no	yes	no	no	no
The Slovak National Gallery, Bratislava	www.sng.sk	-	yes	yes	yes	approx.200 in Slovak version	yes
Russia							
The State Hermitage, St. Petersburg	www.hermitage.ru	-	yes	yes	yes	hundreds	no
The Tretyakov Gallery, Moscow	www.tretyakov.ru	-	no	yes	yes	approx. 600	yes
The Art History Museum, Moscow	www.museum.ru/gmii/	-	no	yes	yes	approx. 700	yes
The State Museum of Ethnography, St. Petersburg	www.ethnomuseum.ru	-	no	yes	yes	hundreds	no
The Palace-Museum, Pavlovsk	www.pavlovskart.spb.ru	-	yes	yes	no	no	no
The State Russian Museum, St. Petersburg	www.rusmuseum.ru	-	yes	yes	yes	approx. 300	no
The Moscow Kremlin Museums	www.kreml.ru	Germ, Span	no	no	yes	dozens	no
The Kunstkamera (Anthropology and Ethnography), St. Petersburg	www.kunstkamera.ru	-	no	yes	yes	approx. 100	no
The Regional History Museum, Omsk	museum.omskelecom.ru/ogik/	Russian only	no	yes	no	no	no
The Art and History Museum, Vladimir	www.museum.vladimir.ru	-	no	yes	yes	approx. 100	no
Asia							
The Israel Museum, Jerusalem	www.imj.org.il	Arab	yes	selected	no	a dozen	yes
The National Museum of India, New Delhi	www.nationalmuseumindia.org	-	no	yes	yes	100	no
Reza Abbasi Art Museum, Tehran	www.rezaabbasimuseum.org	-	yes	yes	yes	600	no
The National Museum, Ulaanbaatar	www.nationalmuseum.mn	-	no	no	yes	dozens	no
The Shanghai Museum	www.shanghaimuseum.net	-	no	yes	yes	dozens	yes
The Singapore History Museum	www.museum.org.sg	-	yes	yes	no		no
The National Palace Museum, Taipei	www.npm.gov.tw	-	no	almost no	yes	a dozen	no
The National Museum of Korea, Seoul	www.museum.go.kr	Jap, Chin	no	no	no		no
Hong Kong Heritage Museum	www.lcsd.gov.hk/CE/Museum/History/	-	yes	no	yes	approx. 1200	no
The National Museum. Tokyo	www.tnm.go.jp	-	no	yes	yes	approx. 400	no

The National Museum, Nara	www.narahaku.go.jp	-	no	yes	yes	hundreds	database is in Japanese
Australia and the Pacific							
The National Museum of Australia, Canberra	www.nma.gov.au	-	yes	no	yes	less than 50	yes
Australian National Maritime Museum, Sydney	www.anmm.gov.au	-	no	yes	yes	less than 50	yes
The National Gallery of Australia, Canberra	www.nga.gov.au	-	yes	yes	no	dozens	no
The Museum of Victoria, Melbourne	www.mov.vic.gov.au	-	yes	yes	no	no	no
The Melbourne Museum	melbourne.museum.vic.gov.au	-	yes	yes	no	no	no
The Berndt Museum of Anthropology, Perth	www.berndt.uwa.edu.au	-	yes	yes	database without images	not available	no
The Museum of New Zealand Te Papa Tongarewa, Wellington	www.tepapa.govt.nz	-	yes	yes	no	no	no
The Aukland Museum	www.akmuseum.org.nz	-	yes	yes	no	no	no
The Anthropology Museum, Hanga Roa, Easter Island	www.museorapanui.cl	Span	yes	yes	no	no	no
The Fiji Museum	www.fijimuseum.org.fj	-	yes	no	no	no	no
South and Central America							
The National Museum of Applied Arts, Buenos Aires	www.mnad.org	Spanish only	no	yes	yes	approx. 40	no
The National Museum of Brazil, Rio de Janeiro	www.museunacional.ufrj.br	-	no	yes	yes	dozens	yes
The National Museum of Colombia, Bogota	www.museonacional.gov.co	Spanish only	no	yes	yes	approx. 60	yes
The Museo del Oro, Bogota	www.banrep.gov.co/museo/home4.htm	Spanish only	no	yes	yes	approx. 100	yes
The Museo Templo Mayor, Mexico City	www.conaculta.gob.mx/templomayor/	Spanish only	no	yes	no	-	no
The National Palace Museum, Mexico City	www.shcp.gob.mx/museo_palacionacional/	-	no	very brief	no	-	no
The National Museum of Fine Arts, Havana	www.museonacional.cult.cu	Spanish only	no	yes	no	a dozen	no
The Museums of the Central Bank of Costa Rica, San Jose	www.museosdelbancocentral.org	-	yes	very brief	no	-	no
The Chilean Museum of Pre-Columbian Arts, Santiago	www.precolombino.cl	Spanish only	no	yes	only illustration of themes	dozens	no
The Turcs and Caicos National Museum	www.tcmuseum.org	-	no	no	no	a dozen	no
Africa							
The Egyptian Museum, Cairo	www.emuseum.gov.*e.g.*	-	yes	yes	yes	approx. 100	yes
The Coptic Museum, Cairo	www.copticmuseum.gov.*e.g.*	French	no	no	yes	approx. 250	yes
The Graeco-Roman Museum, Cairo	www.grm.gov.*e.g.*	-	no	yes	yes	approx. 100	yes

The National Museum of Mali, Bamako	w3.culture.gov.ml/musee/	French only	no	yes	no	no	no
The History Museum, Abomey, Republic of Benin	www.epa-prema.net/abomey/	French only	no	yes	no	no	no
The National Museum of Ghana, Accra	www.geocities.com/gmmbacc/	-	no	yes	no	no	no
The National Museums of Kenya, Nairobi	www.museums.or.ke	-	yes	yes	no	a dozen	no
The National Museum of Namibia, Windhoek	www.natmus.cul.na	-	no	yes	no	no	no
The National Museums, Cape Town	www.museums.org.za/iziko/		yes	yes	no	no	yes
The Pretoria Art Museum	www.pretoriaartmuseum.co.za		no	no	no	no	no

National digital repository for cultural heritage institutions

*Astrid Verheusen**

This paper will address the problem of long-term preservation and permanent access to digitised material of cultural heritage institutions. It starts off with a description of the issues concerning the preservation of digitised material and the way the National Library of the Netherlands addresses these. This will be followed by an introduction to the e-Depot of the National Library, the system in which born-digital material of publishers is stored. Finally, the paper will yield a description of a project with the goal to set up a national digital repository for cultural heritage institutions in the Netherlands.

Over the last ten years Dutch cultural heritage institutions have digitised (part of) their collections on a large scale. The main aim is to increase accessibility to their cultural heritage through availability on the Internet. The National Digitisation Programme of the National Library – *The Memory of the Netherlands* – greatly stimulated this process. In the last four years the programme digitised over 40 collections of 40 institutes. Most of the projects focus on the digitisation of visual attractive images, although an increasing amount of text material is being digitised as well.

Due to this progress in the field of digitisation, experience has grown and a profound knowledge about best practices has been developed. Certainly, there still remain many unsolved questions about selection, copyright, usability and so on, but digitisation as such is no longer a problem. There are, however, growing concerns about the storage and permanent access of the products of digitisation projects. Mostly high quality TIFF-files (Tagged Image File Format) are being created to serve as master files. The majority of the cultural heritage institutions have no well-formulated policy to archive these master files of digitisation projects. Often the focus lies on producing a web site to make images available. Thousands of tiff-files are stored on CD, DVD, tape or hard disk, lacking good administration. Institutions are not aware of the problems of digital preservation. Storage media will deteriorate, data formats and software to render them will become obsolete. If no action is taken, digitised material will not be accessible in the near future.

Although originals can be digitised again in the future, the high costs of scanning and the growing need to reuse the tiffs for new purposes means that this is not a preferable situation. Furthermore, institutions are starting to show a tendency towards digitising material for the purpose of conservation of the original.[1] In this case, the digitised object is a substitute for the original object. Given the high cost of digitisation and the increasing need to reuse digitised material, the government is stressing the need to make sure the expensive digitised material can be used again when needed in the future and therefore pushes towards establishing standards to safeguard long-term preservation of the digital objects.

The problem of storage and permanent access is not limited to digitised material. Over the last ten years, publishers have published increasingly in electronic formats and have even stopped publishing some publications on paper. The deposit task for publications is the core business for every national library. The National library of the Netherlands also has to acquire, register, preserve and give access to every printed publication that is published in the Netherlands. About ten years ago the library became aware that publishers started to move towards electronic publications. In 1994 it was decided to extend the deposit task of the library to electronic publications. Policies and processes had to change because of that. At that time, a system for safe storage of digital publications could not be bought from the shelves. In 1999 a European tender was published to find a technology partner that would be able to develop a system. A durable storage system was requested with specific loading facili-

* Project leader at the Research & Development Department of National Library of the Netherlands, The Hague

1 See for example www.arl.org./preserv/digit_final.html.

ties, based on the Open Archival Information System (OAIS) model.[2] IBM turned out to be the best partner. The system was delivered in December 2002 under the commercial name of DIAS, Digital Information Archiving System. The implementation of the system at the National Library of the Netherlands, where it is called the e-Depot, started in 2003. The e-Depot is now in operational use and regulates the workflow from acquisition to delivery on site by the Acquisitions and Processing Division. From 2002 onwards, archiving agreements have been signed with major publishing companies like Elsevier, Kluwer Academic, BioMed Central, Blackwell Publishing, Oxford University Press and Taylor and Francis.

With the implementation of the e-Depot, safe storage of electronic publications has been made possible. However, a distinction should be made between the safe storage of the bits and bytes and the technology needed to guarantee permanent access to the information itself. File formats, operating systems, software and hardware become obsolete and solutions have to be found to render the information that is being stored. To create permanent access different strategies like emulations, migration and generic viewers are available. At the National Library of the Netherlands large efforts are being made by the Digital Preservation Department within the Research & Development Division to conduct research into each of these strategies.

With the e-Depot in place it is possible to add new services to the system and the organisational infrastructure. The first of these new services was initiated in 2003 with the establishment of the DARE project (Digital Academic Repositories). The DARE project aims to store Dutch academic information of all Dutch universities in the e-Depot. Because of the National Library's own problem with the long-term storage of digitised master-files, the Library also took the initiative to start a project to investigate the possibilities for long-term storage of tiff-files from cultural institutions. The library consulted several cultural heritage institutions about the desirability to build a central national service for this purpose. Their reactions were positive. At the end of 2003 the Ministry of Economic Affairs funded a pilot-project to explore the possibilities for a 'tiff-archive'. The project started in 2004.

Two main objectives have been formulated: the development of a pilot system for the storage of tiff-files and the formulation of a business plan to assess the feasibility of a national service for storage. The pilot is a co-operation of five cultural heritage institutions who help to compose an inventory of the wishes and demands cultural institutions have for such a service. Different kinds of institutions have been invited to join the project in order to be able to test different types of material. The service will be built according to a business-to-business model. In this perspective the project is innovative in the world of the cultural heritage institutions. At the end of the pilot an evaluation will take place and a decision will be made about whether or not the system will be taken into production.

The tiff-archive will make use of the e-Depot system. IBM and the National Library of the Netherlands are building new functionalities for this purpose. The system will have a secured web interface. This enables the institutions to load their material into the archive themselves. The e-Depot originally was not meant for delivering the files back to the publishers. In the tiff-archive institutions can search for there own tiffs en have them delivered back online so they can use the images at their disposal. Customers may query the online catalogue after a process of identification, authentication and authorisation. This way files will only be available to the owners of the collections.

Initially, the digital repository focused on images, more precisely on the master files of digitisation projects. The main problem with tiff-files is their potential size. Tiff-files vary in size from several megabytes to four gigabytes. In this perspective the tiff-archive also differs from the e-Depot: electronic publications mostly are not larger than about one megabyte. In testing the pilot system, the focus will therefore lie on testing the performance of the system when files (of large sizes) are loaded, stored, retrieved and delivered back to the institutions. If the system proves to work for tiffs, other file formats will be admitted.

The project has conducted research into which metadata are needed for the long-term accessibility of tiffs. Until now metadata for digitised images has focused on defining descriptive elements for discovery and identification. Little attention has been paid to defining the metadata that describe the capture proc-

2 See ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf

ess and technical characteristics of the digital images. Technical metadata are much needed as an essential component for short-term and long-term management purposes. For this purpose a profound analysis was made of the draft NISO standard Z3987: Technical Metadata for Digital Still Images.[3] This standard contains over 100 elements, all meant to ensure that images will be rendered accurately in the future and can be managed properly. The standard is very useful for digitisation projects still to be carried out, but proved very complicated to implement for images that already exist. For the tiff-archive a data model was developed that can incorporate the NISO metadata if the institutions add them. In case they have not added any technical metadata, a selection of basic elements must be added to ensure longevity. By selecting only the most important technical elements, the tiff-archive data model is kept practical for use by the cultural heritage institutions.

Because the tiff-archive makes use of the e-Depot system, it can profit from the services that are being developed for long-term preservation and from research into strategies for emulation and migration. One of these research projects is called the *Preservation Manager*. The Preservation Manager will store information on the file formats that are kept in the e-Depot to ensure the rendering of stored files.[4] As long as there is a link in the metadata from a saved object to the Preservation Manager there will be a warning when the file format becomes obsolete.

The tiff-archive will be 'not for profit'. Because the tiff-archive will be built according to a business-to-business model, a business plan is written. Among other subjects the business plan will focus on target customer groups, the user needs, risks (technical and otherwise), costs and the price level of the service. The business plan needs to give insight into the financial needs to realize the service. It should provide insight in up scaling of the pilot to an operational service and the costs of maintenance of the system.

In March 2005 an enquiry was set up to find out if there is a market for a national digital repository and what the demands of potential customers are. This market research was necessary to provide information for the business plan. The enquiry was sent to 300 archives, libraries, museums and documentation centres. Approximately 100 institutions responded. The response did not show any significant differences between the various types of institutions: the answers from archives, libraries, museums and documentation centres were all very similar.

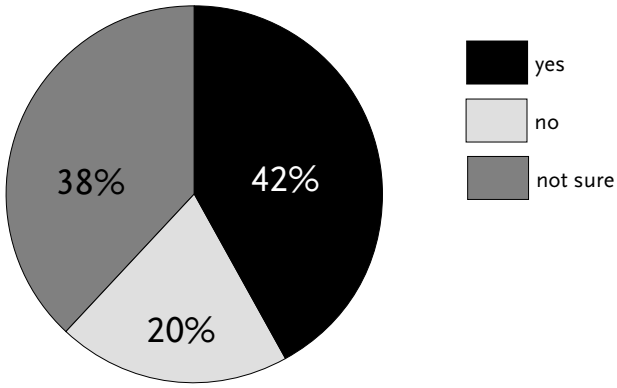
From the 100 responding institutions 74 percent is in possession of tiff-files. 30 Percent holds more than 10.000 tiff-files. The storage of the tiff-files in the various institutions takes up between 100 gigabyte to several terabyte. On the question on which medium the tiff-files are stored, CD-ROM, DVD and hard disks were the most popular. About 57 percent of the institutions indicated that they did not have a policy towards save storage of the tiff-files, while 43 percent said they did. Almost all institutions had plans for digitising more material in the near future.

The question was asked if the institutions would make use of a central repository for digitised material. The figure below shows the result of the answers.

When asked for reasons why the institutions would not make use of a central tiff-archive, the institutions answered they did not see the need for a central repository because the storage of tiff-files is so easy they can manage themselves, because they use the files often and think a repository will not be able to deliver the files quickly, or because they do not outsource the storage of their paper files so why outsource the storage of their electronic files? Institutions that were not sure of the need for a central repository had many questions about the service. They wanted to know how fast the performance would be, had questions about conditions and had doubts about copyright, authenticity, accessibility, the National Library as institution that would deliver such a service and most of all about the costs of the service.

The response of the institutions shows that many of them are not always aware of the problems concerning safe storage and permanent access. Sometimes institutions confused simple storage with long-term storage and digital preservation. The institutions that do see the use of a central repository are very demanding towards the performance of the service. Lastly, nearly all institutions worry about the costs. However, an astonishing 62 percent of the institutions that did see

3 See www.niso.org/standards/resources/Z39_87_trial_use.pdf
4 For more information see: www.kb.nl/hrd/dd/dd_onderzoek/preservation_subsystem-en.html



Need for a central repository for digitised heritage material

the need for a central service were aware of the high costs and were willing to pay.

A lot has been written about the costs of digital storage and preservation. It is not easy to find a standard model that covers all relevant expenses to calculate the costs. A first inquiry to calculate these at the National Library was done in the beginning of 2005. The model that was developed is based on the costs that were made for the storage of two million articles in the e-Depot. Personnel, hardware and maintenance costs of storage in the e-Depot were taken into account. Costs for the development of strategies for preservation were not included. The outcome shows that safe storage is not cheap. Storage of one gigabyte costs about € 160 per year. This means that the storage of a small collection taking up about 100 gigabyte would cost € 16.000 per year, which is not affordable for most institutions. The figures are still very premature and have to be specified further in the business plan, but cost will definitely be high.

At the end of 2005 the pilot system for a national digital repository is expected to be operational. Then the technological problems and possibilities of a tiff-archive will become clear. Furthermore, the pilot system will provide further insight into the costs of building and maintaining a national service for this purpose and about the wishes and demands of the cultural institutions. As the enquiry above shows, cultural heritage institutions need to become more aware of the problems concerning digital preservation of their digitised material. Guidelines en recommendations will be drafted to inform them about the best ways to digitise en store their digitised collections. A combined effort to tackle the challenges in this area is one of the most critical success factors for future access to our cultural heritage.

Virtual libraries and thematic gateways in German history: Strategies and perspectives

*Max Voegler**

Introduction

A recent strategic report commissioned by the German Federal Ministry for Education and Research (BMBF), entitled 'Networking Information – Revitalizing Knowledge' ('Information vernetzen – Wissen aktivieren'), begins with a vision:

A scientist enters her office in the morning. She calls up the internal database of the geographic information system of the current experimental setup as a flow chart on the panel display on the laboratory wall. While she readies her work-space for the day, the system searches in external, internal and commercial databases, whether any new research exists regarding the various sub-fields of her current work. A certain reaction cycle, which preoccupied the Researcher on the way to work, is drawn on the board, which the system automatically adds to its search query. Some time later it presents several results from quality peer-reviewed journals and databases in the field of chemistry. The system then proceeds to read the titles of the results in an intelligent ranking based on the search queries originally entered. Using her voice, the researcher can now tell the computer which publications to add to her personal library and which to throw away.[1]

Such a 'vision' – here skewed toward the natural sciences – could easily be applied to historians as well: entering her office in the morning (or, better yet, while making her morning cup of coffee at home), the historian immediately sees the newest websites and projects, publications, calls for papers, conferences and so on related to her research and teaching interests that were found by intelligent agents within the past day, week or month. After reviewing the various resources, downloading any interesting papers for reading, maybe ordering a book or two mentioned in a new review, and adding a conference and call for paper or two in her calendar, the historian is able to instantly pull together a set of electronic resources on a new topic she has been working on.

Recent developments in technology have made that vision, though not yet a reality, then something that at least can almost be imagined. Vast depositories of secondary literature and of digitalized primary sources, myriad internet sources cataloged along library classification systems, intelligent agents and meta-search engines, and, of course, virtual libraries and thematic gateways to give the information overload a sense of order, and to deliver the relevant results to your doorstep.

Concepts in virtual libraries

Virtual libraries are places in which scholars find the information they need for their research and teaching in an electronic form. This includes elements of electronic libraries, or 'libraries without walls,' – an 'aggregation of catalogs, lists and indexes of documents of every imaginable type, organized according to myriad

**Humboldt-Universität zu Berlin*

1 Bundesministerium für Bildung und Forschung, *Information Vernetzen – Wissen Aktivieren. Strategisches Positionspapier des Bundesministeriums Für Bildung und Forschung Zur Zukunft der Wissenschaftlichen Information in Deutschland* (Bundesministerium für Bildung und Forschung, 2002 [cited 10 November 2004]); available from http://www.bmbf.de/pub/information_ vernetzen-wissen_aktivieren.pdf.

Figure 1. Search for 'Napoleon' in the Clio-online Web-Directory. (20 April 2005)

schemes of classification;'[2] but it also goes beyond that definition. Virtual libraries, especially when they are subject-based, act both as large-scale repositories and as filters, seeking to combine *relevant* available online resources and to present that information in a manner transparent to scholars within that discipline. On one level, this implies that more than just traditional texts are included: next to books and articles, historians would expect also to find digitalized versions of primary sources; conferences (dates and deadlines), calls for papers and other forms of information that are important to the discipline but outside of the 'traditional' scope of libraries. On another level, it also implies less than a traditional library: the information must also be filtered – much of it excluded – and presented in a logical manner.

In this section I will focus on three approaches within virtual libraries that seem to me to be most important: browsing, searching, and personalizing information. To begin with ***Browsing***: Portal or gateway sites order the information they contain for the user and this means providing a transparent browsing structure. As has often been noted, browsing structures have inherent drawbacks, 'whether the user is browsing or searching, whether they are an eighth grade student or a Nobel prize winner, the identical information is selected and it is presented the same way.'[3] Indeed, the reason that there are virtual libraries and thematic gateways to any number of disciplines and sub-disciplines is that each discipline classifies information in different ways, has different priorities, and collects different sorts of data. And while the first-time user

2 Geoffrey Nunberg, 'The Places of Books in the Age of Electronic Reproduction,' *Representations* 42 (1993): p. 30.
3 Susan Gauch, Jason Chaffee, and Alexander Pretschner, 'Ontology-Based Personalized Search and Browsing,' *Web Intelligeant and Agent System* 1, no. 3-4 (2003): p. 1.

An der Schwelle zur Moderne : Deutschland um 1800 [Historisches Forschungszentrum der Friedrich-Ebert-Stiftung]

Description	"Die digitale Bibliothek der Friedrich Ebert Stiftung ermöglicht den umfangreichen Zugriff auf Aufsätze verschiedener Autoren über die Zeit um 1800, die Ära der Befreiungskriege und darüber hinaus: vom Mythos um die preußische Königin Luise bis zur Verfassungsfrage, sozialen Lage und zur Situation der Pressefreiheit. Äußerst übersichtlich und klar strukturiert wie ein digitales Lesebuch zu lesen, downzuloaden und für die eigenen Forschungen sehr gut zu gebrauchen." Copyright: Virtuelles Museum Preussen <http://www.virtuelles-museum-preussen.de>
URL	http://www.fes.de/fulltext/historiker/00671toc.htm
Publisher	Brandt, Peter
Published by	Friedrich Ebert Stiftung - Digitale Bibliothek
Country	Deutschland
Language	Deutsch
ISBN	3-86077-863-3
Reviewed	7/22/2005
Add. Information	
German Keywords	Politische Entwicklung, Sozialer Wandel, Verfassung, Krieg, Befreiungskriege, 1813-1815, Widerstand, Presse, Luise, Königin von Preussen, Napoleon I. Bonaparte, Kaiser d. Franzosen
English Keywords	political development, social change, constitution, war, liberation wars, resistance, press
Tech. Provider	Friedrich Ebert Stiftung; Bonn, DE <http://www.fes.de>
Derived from	An der Schwelle zur Moderne : Deutschland um 1800 / Peter Brandt (Hg.). Forschungsinstitut der Friedrich-Ebert-Stiftung, Historisches Forschungszentrum. - [Electronic ed.]. - Bonn, 1999. - 183 S. = 490 Kb, Text . - (Gesprächskreis Geschichte ; 31).
Access	free
Restriction	none
Datasource	Clio-online

Figure 2. Example of fields within an entry. Clio-online Web-Directory. (20 April 2005)

will more often than not ignore browsing structures altogether, they become increasingly important to returning users.

Alas, in the age of Google, users tend to ***search*** more than they browse. And, judging not just from my own behavior,[4] users first search for some aspect of information that they need and then, through that snippet, discover the rest of the sites. This calls for very different sorts of search strategies: Firstly, when a user initially explores a new sight, he or she tends to enter the search term of primary interest at that moment (*e.g.*, 'Napoleon' as an example, or 'digital sources'), without thinking about which categories or types of resources he or she is expecting to retrieve. If the site is to become useful to the user, its designers need to use the search process as a method of teaching the user about the site – of bringing the central categories and concepts in use to his or her attention. And this should happen in the search results window, where the categories of each entry are displayed as part of the results (See Figure 1); and it needs to happen in the individual entries, where keywords associated with that specific entry help the user discover other, similar results. (See Figure 2).

Secondly, a site needs to enable a user to carry out complex searches within a given set of results using what is called a spider or dynamic agent. Spiders catalog and index pre-specified sub-sets of sites, thus enabling the user to search not just within the internet portal itself but within a specified subset of resources contained within the gateway.[5] On a simple level, this means that one could create a content index of all

[4] See, for example, Jakob Nielsen, *Is Navigation Useful?* (Januar 2000 [cited 11. Mai 2005]); available from http://useit.com/alertbox/20000109.html

[5] On spiders and web-mining, see, Hsinchun Chen, 'Editorial Special Issue Web Retrieval and Mining,' *Elsevier Decision Support Systems*, no. 35 (2003); Hsinchun Chen et al., 'A Smart Itsy Bitsy Spider for the Web,' *Journal of the American Society for Information Science* 49, no. 7 (1998).

terms in the web sites within the browsing category 'archives,' which, returning to the example of 'Napoleon,' would enable the user to search for archival materials on Napoleon within all resources on archives contained within the portal. In a more complex example, it means that a user can dynamically create a set of sites to be contained in an index and then continue searching within that set, i.e., all pages found in the initial 'Napoleon' search or a suggested subset of pages that are based on a selection of resources that were employed by previous users using the same or similar search queries and key words.

Finally, a third search strategy is based around so-called meta-search engines, which enable searching multiple databases, library catalogs and other resources simultaneously. Here, it should be noted right off the bat, much depends on the institutional and national context with which a user accesses information. Though within the North American context much of the bundling of disparate electronic resources – at least in the pay-per-view sector – has been handled by private companies such as Chadwyck-Healy, within the German and European context, many of the individual resources, especially databases, are produced using public research money and thus often freely available in a central gateway search. Alas, these are not readily accessible, searchable or findable using standard search engines such as Google, MSN or Yahoo. A meta-search engine for historians thus combines the relevant available databases in a central format, though with the growing number of databases available, this selection also needs to be customizable: only archival resources, or library catalogs, etc.

The different browsing and searching strategies lead to a logical conclusion: **personalization** must become a core component of large information depositories. Related to browsing, this is a simple task: once a user has entered his or her preferences (*e.g.*, 'Napoleon,' 'Nineteenth Century,' 'Military History') and these have been translated into the classification scheme in use in a web site, new entries, updates or newly found resources can be indexed and presented to the user dynamically. Here we finally get closer to the vision of the historian presented at the beginning of this paper. Finally, applying personalization strategies to searching creates a further possibility. Comparing search words with further search words entered for a previous search; 'remembering' the categories last browsed; or by using the user profile – all these methods can help significantly narrow a search at the outset.

German history online

Needless to say, we still have a long way to go. But what I have outlined above is, I think, the general vision we should keep in our heads as we think about thematic gateways and virtual libraries. In this next section I will look at the current state of online-resources available within German history, examining several different projects and then presenting a possible roadmap on the basis of the project at which the author works.

In a recent paper on the state of historical studies online in Germany, Ewald Grothe listed the six sites he found central for historical scholarship online.[6] Next to Clio-online Grothe went on to list three projects associated with Clio: *H-Soz-u-Kult*,[7] a mailing list that is part of the H-Net; *Zeitgeschichte-online*,[8] a site devoted to contemporary history that is technically based on Clio-online and run editorially by the Center for Contemporary History in Potsdam; and the *Jahresberichte für Deutsche Geschichte* (Annual Reports on German History)[9] and so-called long-term project based at the Berlin-Brandenburg Academy of Sciences. As well as two further projects: *historicum.net*,[10] a gateway site similar to Clio-online but with its roots firmly in the early-modern period; and the *Nachrichtendienst für Historiker* (Newsservice for Historians), an ambitious site devoted to newspaper links, forums, book publications, and other technical services, not just for historians but also for the broader public.[11]

While I certainly can recommend the last two sites – historicum.net and the NFH – both of which are not directly connected with my own, project, Clio-online, I

6 Ewald Grothe, 'Geschichte im Netz. Ausgewählte Fachinformationsmittel der Geschichtswissenschaft im World Wide Web' (Hausarbeit zur Lehrveranstaltung 'Informationsmittel und -ressourcen' von Prof. Hermann Rösch, Fachhochschule Köln, 2004).
7 See, http://hsozkult.geschichte.hu-berlin.de/
8 See, http://www.zeitgeschichte-online.de/
9 See, http://jdg.bbaw.de/cgi-bin/jdg
10 See, http://www.historicum.net/
11 See, http://www.nfhdata.de/premium/index.shtml

want to focus on Clio-online and the other sites to look at how different types of content, different search and filter strategies, and different 'audiences' can emerge from bundling information across portals. Each portal has become successful because it has found a certain niche within the discipline:

Clio-online acts as a central gateway to historical resources. It does not 'produce' editorial content for its site and instead serves as a conduit, presenting the landscape of history online to the user and enabling him or her to find the resources he or she is looking for efficiently. It does produce informational guides to various topics as well as a series of e-publications in cooperation with H-Soz-u-Kult as well as other partners. It is funded by the Deutsche Forschungsgemeinschaft (DFG) within the funding framework 'scientific access to literature and information' ('Wissenschaftliche Literaturversorgungs- und Informationssysteme') and thus also follows the general aims of the program, which are to provide a central platform within the discipline for accessing scholarly information.[12] Its role is thus that of the 'meta'portal: pulling together various types of information and presenting and personalizing them for the (academic) user.

H-Soz-u-Kult, on the other hand, *is* content. It has become *the* central information platform for German-speaking historians, with over 10,000 subscribers, 750 book reviews, 250 conference reports, and another 1,000 CFPs, lectures, colloquia and tables of contents annually. H-Soz-u-Kult enjoys some funding through Clio-online (and thus indirectly through the DFG) but mostly rests on the editorial work of over 30 scholarly volunteers and the technical infrastructure of the H-Net staff in Michigan and the Humboldt University in Berlin. H-Soz-u-Kult works with Clio-online to create e-publications, reviews web sites featured in the Clio-online directory, and closely integrates its other content with Clio-online. But H-Soz-u-Kult – and this is important to the scholarly community that has made it its home – remains editorially independent from Clio-online and the other institutions involved in that project.

Zeitgeschichte-online (ZOL) is largely a sub-component of Clio-online: It is built on the same software platform and uses the same relational database as Clio-online, but limits the resources it displays to a subset devoted to contemporary history. Furthermore, this ambitious project also has an e-publishing component, the periodical *Zeithistorische Forschungen* as well as such services, as well as a press digest and television schedule on contemporary history. Finally, ZOL has taken over the editorial responsibilities for texts relating to contemporary history within H-Soz-u-Kult. It is thus part H-Soz-u-Kult, part Clio-online, and wholly a subsection of both, based on filtering out relevant content from other connected projects for a specific sub-discipline within history.

The *Jahresberichte für deutsche Geschichte* (JDG) provide an online bibliography of current research related to German history in all publications. All its entries from 1985 onward (currently over 230,000) are online and freely available; they are also fully cataloged and classified by keywords and time period. As a wonderful but specialized type of resource, however, the JDG are hardly known outside of a select circle of what one may term 'internet insiders.' But in partnership with Clio-online and ZOL, for example, it has begun producing subsets of its database – for example related to the First World War or to contemporary history – which it then integrates into the other portals' search strategies.

These projects are thus not isolated endeavors; they are interconnected in a complex web of services. To return to the themes of browsing, searching and personalizing information presented above: Clio-online acts as a central aggregator within the discipline, collecting and combining different types of resources. Within the realm of cataloging internet resources, for example, we work together with a consortium of libraries and online projects[13] to create a central pool of resources from which each project can then take those resources it needs for its own sub-discipline. Each collaborating project then uses its own sets of filters to generate a dynamic subset of that pool for its own virtual library.

12 The DFG allows projects to apply for start-up financing (it will usually fund a five-year period, though a project has to re-apply for funding at least every two years) with the assumption that, after this period, a successful project will be taken over by the institutions that applied for the funding.

13 On the so-called 'Network Subject Gateways History,' http://www.historyguide.de/varia/netzwerk/websitesfuerhistoriker.pdf

These can then be classified and presented in the manner most fitting for the purpose. At present, the partners have mostly a regional focus – Eastern Europe or Latin America – but this will undoubtedly change in the future as other institutions join the project. The manner in which these final projects are intertwined shows the direction toward which historical online projects are moving in Germany: not a consolidation into a single project, but a set of thematic and specialized web portals and forums that are increasingly interlinked through common standards and meanwhile maintaining disparate goals.

References

Bundesministerium für Bildung und Forschung. 2002. Information Vernetzen – Wissen Aktivieren. Strategisches Positionspapier des Bundesministeriums Für Bildung und Forschung Zur Zukunft der Wissenschaftlichen Information in Deutschland. In, Bundesministerium für Bildung und Forschung, http://www.bmbf.de/pub/information_vernetzen-wissen_aktivieren.pdf. (accessed 10 November, 2004).

Chen, Hsinchun. 'Editorial Special Issue Web Retrieval and Mining.' *Elsevier Decision Support Systems*, no. 35 (2003): 1-5.

Chen, Hsinchun, Yi-Ming Chung, Marshall Ramsey, and Christopher C. Yang. 'A Smart Itsy Bitsy Spider for the Web.' *Journal of the American Society for Information Science* 49, no. 7 (1998): 604–18.

Gauch, Susan, Jason Chaffee, and Alexander Pretschner. 'Ontology-Based Personalized Search and Browsing.' *Web Intelligeant and Agent System* 1, no. 3-4 (2003): 219-34.

Grothe, Ewald. 'Geschichte im Netz. Ausgewählte Fachinformationsmittel der Geschichtswissenschaft im World Wide Web.' Hausarbeit zur Lehrveranstaltung 'Informationsmittel und -ressourcen' von Prof. Hermann Rösch, Fachhochschule Köln, 2004.

Nielsen, Jakob. 2000. Is Navigation Useful? In, http://useit.com/alertbox/20000109.html. (accessed 11. Mai, 2005).

Nunberg, Geoffrey. 'The Places of Books in the Age of Electronic Reproduction' *Representations* 42 (1993): 13-37.

A new approach: The arrival of informational history

*Toni Weller**

In recent years there has been an emerging trend of academics from both disciplines of History and Information Science looking at what are fundamentally similar concepts. However, these have been from such different approaches that we have been left with a 'grey area' of missed opportunity for research. The author of this paper has a background in both disciplines, and will attempt a start at redressing the balance by examining the historiography of the topic and exploring the potential of 'Informational History'. It will conclude with a look at the interdisciplinary future for History and Information Science.[1]

Previous accounts by Historians and Information Scientists have been limited by their focus on their own disciplines. Many writings on the history of Information Science could be criticised as being ahistorical for ignoring the context of time, and the different ways in which things were viewed by contemporaries (explored more fully below) – a fundamental issue for the Historian. Often these 'histories' of Information Science focus on the development of a particular technology, or the rise of the internet, or place modern day assumptions on why and how things developed as they did – a dangerous teleological position in itself. Criticisms can also be made of Historians who write on Information Science themes. History as a discipline is centred around evidence and information – how it is used, by whom it is used, and what can be discovered from all aspects of it. Yet historical accounts tend to skim over concepts such as 'the information society', the commoditization of information, and most significantly, the role and impact of the actual information *itself*, all of which have their own literatures within the Information Science discipline.

This is not to say that the existing research has no value or is erroneous, but rather that there is potential for a revision and challenge to accepted views of both Historians and Information Scientists. There is a new discipline emerging in the form of Informational History, which allows for a deeper understanding of the role information and its uses (and misuses) has played in the past, and thereby enriching both of its parent disciplines. History is, after all, a dynamic, changing story. One of its key purposes is to provide understanding of change, and Information Science as a discipline is often regarded as being a manifestation of a rapidly changing society (Saracevic, 1999; Warner, 2000). Such a combination should also allow for a deeper discussion of the issues central in our own culture (the information society, digital divide, information literacy, privacy rights, etc) enabling us to place them in their wider historical context, and flesh out a more three dimensional picture of the thematic roots of Information Science. This paper is not proposing however, that Informational History is an umbrella term for the history of literacy, or the history of the press, or the history of privacy and so on. Rather, it argues that Informational History is a way of looking at established historical themes by attempting to tease out how they affected people, social and political values, and cultural norms, within an *informational* context. In other words, focusing on information as a

*Department of Information Science, City University, Northampton Square, London, EC1V OHB, t.d.weller@city.ac.uk

1 This paper is based on preliminary PhD research considering the impact of the commoditization of information in England in the latter half of the nineteenth century, and has its foundations in an MA in History and an MSc in Information Science. Although this paper focuses on the nineteenth century, it is argued that an Informational History approach could (and should) be applied to any period.

central theme, and examining its impact upon existing historical theses, and contemporary social, cultural, or political infrastructures. The point is not to compare historical periods but to focus completely upon one era and gain a full sense of the informational issues as well as the historically traditional areas of politics, economics, culture and society.

There have of course been attempts at interdisciplinary work in this area, both in terms of analyses by Historians which consider themes central to Information Science, and 'historical' works by Information Scientists. Information Scientists have been using variations of the terms 'historical information science', or 'historical informatics' for the last decade or so (Marvin, 1987; Karvalics, 1994; Warner, 2000; Karvalics, 2003). Similarly, there have been Historians who have focused on the history of science and technology, or the rise of the 'information state' (Eastwood, 1989; Agar, 2003; Flichy, 2004; Higgs, 2004). However each of these writings has ultimately been focused upon its own discipline, and has to a large extent neglected the methodologies and ideologies of the other, thereby overlooking trends and similarities which could provide key opportunities for revisionist study.

Examples of Historians' work on informational themes include Eisenstein's analysis of the printing press as an agent of change (1979), Eastwood's look at the State's use of information in the nineteenth century (1989), Winston's account of the telegraph and telephone (1998), and Robertson's discussion of Hypertext History (2004). These are Historians who, intentionally or not, are embracing some of the concepts of Information Science into their own work; communication, control of information, role of new technologies, and digital media.

These kind of historical accounts tend to fall into two broad categories. First, there are those which have a tendency to consider the impact of communicational and organisational mechanisms and tools, but not of information itself (such as Eisenstein, Eastwood and Winston). These are relevant areas of historical study, but they do not work as examples of Informational History because they take information itself for granted within their discussions, and do not consider its contextual impact.

Second, there are those which consider new ways of teaching, researching, or studying History by utilising new media forms and digital technologies (such as Robertson). Recent work by Information Scientists Boonstra *et al* (2004) covers similar ground as Robertson by looking at the 'past, present and future of historical information science'. Here we see a reoccurring trend as what is actually being discussed is how computer and digital technologies can be used within the Historical community to facilitate research, teaching, and the storage and retrieving of information, or in other words, 'computerised methods to be used in historical science' (2004, p.9). This is undeniably important in deepening relations between Information or Computer Scientists and Historians (and follows the trend of Geographic Information Science), but it not Informational History. It is too focused on the technical, on the practical application of informatics in Historical study, rather than on the historical past itself. However, Boonstra *et al* go to great lengths to explain that 'historical information science is neither 'history' nor 'computing'. It is a science of its own, with its own methodological framework', thereby supporting the argument that there is an emergent trend of interest between both disciplines. The report goes on to state that 'the object of historical information science is historical information, and the various ways to create, design, enrich, edit, retrieve, analyse and present historical information with help of information technology' (p.10). While valuable in its own right, a methodology of this kind is not that of Informational History, which instead attempts to re-interpret the historical past with an informational emphasis.

The nineteenth century has also been written about by Information Scientists when attempting to trace the history of the discipline (Beniger, 1986; Schement *et al*, 1987b; Bawden & Robinson, 2000). However, these works, almost without exception, focus on the development of specific communication technologies and ways of organising information – the telegraph and telephone in particular, and later the radio, typewriter, classification schemes – without considering the differing contexts in which each of these technologies emerged, or their contemporary impact. Some Information Scientists, such as Karvalics (2004) and Warner (2000), use the term 'historical informatics' to describe this focus on the historic development of certain scientific methods and technologies. These accounts have a tendency to be deterministic, and often at least indirectly support the idea of a teleological progression into the 'ultimate' form of information society. These accounts are both misleading and reductionist, ignoring the role and impact of social values

and cultural norms on the development and integration of technologies.

Similar thematic approaches to the history of a particular technology or organisational model have also been made by some Historians (Pool, 1977; Winston, 1998; Agar; 2003, Higgs; 2004). However, these have often been regarded as part of the fringe subject of the history of science and technology, whose focus is consequently not on the informational impact and context. So, Agar's fascinating work on the government machine is written from the perspective of the history of science and technology, rather than with an informational context, as is his research interest. Similarly, Higgs' wide writings on the history of censuses and surveys, civil registration, and the impact of the digital revolution on archives can be seen through his focus on these themes when he writes about the 'information state'. In other words, although both Agar and Higgs are aware of the historical context of the periods they are writing about, they still focus upon the practical technological and organisational issues, rather than examining the contemporary social, cultural, and political impact of such informational themes. What it does show once again, is the emergence of overlap in methodology and content between History and Information Science.

To date, some of the most truly *historical* information science research has been in the form of micro studies on the role of libraries in the nineteenth and early twentieth centuries, which really focus on the specific historical context of these organisations (Black, 2003, 2004a, 2004b). Once again however, the focus tends to be upon the Information Science key themes of organisational tools within libraries, or the history of libraries themselves, rather than the contemporary social or cultural impact of having such information so accessible. Consequently, works such as Black's, while urging the importance of historical study within Information Science, are still limited because of their lack of understanding of how these institutions fit within the wider historical context. Conversely, Eastwood's (1989) interpretation of the changing role of the state in early nineteenth century England shows an awareness of how access and control of information *itself* can influence existing socio-economic relations, a view which is more usually adopted by Information Scientists in light of our contemporary society. There is no reason however, why such research should be so limited within disciplines.

Recently there have been arguments that Information Science needs to embrace more broadly accepted theories from other disciplines and consider the wider theoretical and philosophical basis of its own field in order to progress (Karvalics, 1994; Warner, 2000; Hjorland, 2005). Similarly, while History is a much older and more established discipline than that of Information Science, and certainly could not be criticised of having too narrow a basis, it must continue to embrace new academic theories and disciplines, and to challenge accepted views of the past. Indeed, in Agar's recent work he supports this view, arguing that 'information – what it meant and how it was collected and used – must be understood in terms of its context. Straws in the wind suggest that an informational history is emerging. Historians of an older generation, including Alfred Chandler Jr. and Robert Darnton, are re-emphasizing informational aspects to their own work to reinterpret business and cultural history... Information serves as a way of dressing up studies of postal systems or histories of the Enlightenment publishing, but it also provides a means of bringing troublesome new technology into historical narratives, thus domesticating it. There is potential in a new informational history' (2003, p.13)[2]. Indeed, Agar's own research has tended to focus upon the history of science and technology and *The Government Machine* is consequently a reinterpretation to 'emphasise informational aspects'. Along similar lines, Karvalics (1994) promisingly argues that in 'the present state of the development of the different branches of informatics there is a need to turn toward historical approach'. However, this is then clarified to be a 'study of the history of information systems' (1994, p.1). There remains a focus on new *technologies* within existing 'interdisciplinary' work by Historians and Information Scientists, and it is this technological determinism which Informational History attempts to avoid.

'Information' is undoubtedly an ill-defined word, or perhaps rather the trouble is that it has been too defined. Within Information Science there has been

2 Robert Darnton, 'An early information society: news and media in C18 Paris', in American Historical Review, 105 (2000), pp.1-35.
Alfred D. Chandler and John Cortada, eds. 'A nation transformed by information', Oxford University Press, 2000.

much debate on different levels of 'information', of what constitutes it in the first place, of the difference between gossip, facts, information and knowledge (Braman, 1989; Haywood & Broady, 1994; Rowley, 1998; Bawden, 2001; Boonstra *et al*, 2004, p.11). While such deliberation is useful to an extent, there can be a danger of over analyzing terms, and thereby losing sight of the bigger picture. Defining 'information' in an historical context is even harder because the values contemporaries attached to such words can be so different from those used in the 2000s. Indeed, Information Scientists writing 'historically' have repeatedly attempted to fit modern concepts such as information and document management, the internet, and so forth into the developments and technologies of earlier centuries. Bawden & Robinson's (2000) discussion of the similarities of the communications revolutions of the fifteenth century and the twentieth century (with the introduction of the printing press and the internet, respectively) is one such example, although the authors do acknowledge the limitations of the study. It could be argued from an Information Science viewpoint that this is a way of writing about the past in terms of current concepts. However from an historical point of view, this is unlikely to be viewed as 'good' history because the current concepts dominate to the detriment of the historical context. Informational History would not attempt to view the past in these terms, but rather view the historical context as it was, but focus on informational themes within this context, as opposed to established themes such as age or gender.

Karvalics (2003) examines the history of the internet, and in doing so constantly jumps from the 16th century to the 21st century in a way which not only suggests an inevitable teleological progression, but which also completely omits any reference to the context in which these technologies were developed, or why and how they were used within society. Informational History should not focus upon the question of what information is, but should perhaps adopt the Popperian approach of defining differing concepts – in this case 'informational' ones – depending upon the context in which they are being viewed (Popper, 1979, esp. p.310). Therefore, to some degree the very flexibility of the word is an asset and should not be forced into a twenty-first century box, as has often been the result with previous Information Science historical approaches. While information is not always necessarily textual, and could certainly also be oral or visual, this definition can be made within each study rather than forcing an explicit definition from the outset, as each will be different. Unlike the concept of information within Information Science, Informational History encourages the exploration of what constitutes information in different historical periods and contexts. Therefore, the Informational History question should not be 'what is information?', but rather, 'what impact does such information have upon contemporary social, political and economic values and behaviour?'

It would be inappropriate to talk so much about context without suggesting why Informational History has appeared at this point in time rather than any other. Everyone is a product of the society in which they exist, and the values which are deemed important by that society cannot fail to influence the way in which individuals, particularly Historians, consider the past. The fact that Information Science appeared as a discipline through the mid-twentieth and early twenty-first centuries is reflective of the growing awareness of technology, networked connections, surveillance, worldwide instantaneous communications, and so forth in our everyday lives. None of these topics are new in themselves, but they are more prevalent in our social values and cultural norms than ever before. Consequently, our approach to the past is inevitably viewed in this context. The implications for both disciplines are significant. Adopting a truly historical methodology to Information Science research lends credence to the discipline academically. Likewise, Historians must recognise that the informational values of twenty-first century society has allowed for new interpretations of the accepted historical past. Both of these are highly significant to our understanding of the world in the early twenty-first century.

In summary, Informational History is neither pure 'History' nor pure 'Information Science', but it does adopt methodologies and key themes from both disciplines. It is 'softer' than traditional Information Science, focusing less on technological and organisational aspects of change and more on the contemporary social, political and cultural impact. It is a reinterpretation of traditional historical opinions, as accepted views are challenged by a re-examination of historical periods in an informational context. Informational History as methodology is not limited to any particular historical period. The informational context will be different in each time period, in each country, for each social class, for differing socio-economic, political,

and cultural interpretations. Consequently, this allows for almost limitless research possibilities since 'information' has been used and misused in some form in every society since prehistory. If Informational History is to reach its potential and acquire the status it deserves, it must reach the standards of excellence of any serious historical investigation, but it must also show an awareness of the twenty-first century in which it is being written. Historians must be objective and try at all costs to understand the period of which they are writing - but as our contemporary values change and evolve they also allow us to revisit our past and shed light on aspects previously overlooked. Movements in this direction by both Historians and Information Scientists would suggest that Informational History has arrived.

References

Agar, Jon (2003), *The government machine: a revolutionary history of the computer*, MIT Press.

Bawden, David; Robinson, Lyn (2000), A *distant mirror?; the internet and the printing press*, in Aslib Proceedings, 52 (2), pp. 51-57.

Bawden, David (2001), *The shifting terminologies of information*, in Aslib Proceedings, 53 (3), pp. 93-98.

Beniger, James R. (1986), *The control revolution: technological and economic origins of the information society*, Harvard University Press.

Black, Alistair (2003), *False optimism: Modernity, class, and the public library in Britain in the 1960s and 1970s*, in Libraries and Culture, 38 (3), pp. 201-213.

Black, Alistair (2004a), *Hidden worlds of the early knowledge economy: libraries in British companies before the middle of the 20th century*, in Journal of Information Science, 30 (5), pp. 418-435.

Black, Alistair (2004b), *National planning for public library service: The work and ideas of Lionel McColvin*, in Library Trends, 52 (4), pp. 902-923.

Boonstra, Onno; Breure, Leen; Doorn, Peter (2004); Past, present and future of historical information science, in *Historical Social Research/Historiche Sozialforschung*, 29 (2), Netherlands Institute for Scientific Information, and the Royal Netherlands Academy of Arts and Sciences, also http://www.niwi.knaw.nl/en/geschiedenis/onderzoek/onderzoeksprojecten/ppf_of_his/final_report/toonplaatje (last visited 01/03/2005).

Braman, Sandra (1989), *Defining Information: an approach for policymakers*, in Telecommunications Policy, 13 (3), pp. 233-242.

Eastwood, David (1989), *Amplifying the province of the legislature: the flow of information and the English state in the early nineteenth century*, in Historical Research, 62 (149), pp. 276-294.

Eisenstein, Elizabeth (1979), *The printing press as an agent of change: communication and culture in early modern Europe*, Cambridge University Press.

Flichy, Patrice (2004), *The imaginary internet: how Utopian fantasy shaped the making of a new information infrastructure*, in Business and Economic History, Vol. 2, pp. 1-1, or available online at (last viewed 28/01/05): http://www.thebhc.org/publications/BEHonline/2004/Flichy.pdf

Haywood, Tim; Broady, Judith (1994), *Macroeconomic change: information and knowledge*, in Journal of Information Science, 20 (6), pp. 377-388.

Higgs, Edward (2004), *The information state in England: the central collection of information on citizens since 1500*, Palgrave Macmillan.

Hjorland, Birger (2005), *Library and information science and the philosophy of science*, in Journal of Documentation, 61 (1), pp. 5-10.

Karvalics, Laszlo (1994), *The claims, pre-history and programme of historical informatics*, in Periodica Polytechnica Ser.Hum.and Soc.Sci., (1), pp.19-30, also http://www.ittk.hu/english/docs/historical_informatics_zkl.pdf (last visited 15/02/2005)

Karvalics, Laszlo (2003), *Internet: the real pre-history and its implications for social theory*, http://aoir.org/members/papers3/Maastrichtprehist_en.pdf (last visited 15/02/2005)

Karvalics, Laszlo (2004), *Information Society Visions: from the early utopies to the adequate government-level strategic planning methods*, in: Informatisation et anticipations. Information Society: Looking ahead Proceedings, June 10-12, pp.63-74, also http://www.creis.sgdg.org/manifs/collocs/is98_actes%20colloque/karvalics.htm (last viewed 15/02/2005)

Marvin, Carolyn (1987), *Information and History*, in The ideology of the information age, Slack, Jennifer and Fejes, Fred (Eds.), pp. 49-62.

Pool, Ithiel de Sola (Ed.) (1977), *The social impact of the telephone*, MIT Press.

Popper, Karl (1979), *Objective Knowledge; an evolutionary approach*, revised edition, Oxford University Press.
Robertson, Stephen (2004), *Doing History in hypertext*, in Journal of the Association for History and Computing, 7 (2), or available online at (last visited 27/10/04): http://mcel.pacificu.edu/JAHC/JAHCVII2/ARTICLES/robertson/robertson.html
Rowley, Jennifer (1998), *What is information?* in Information Services and Use, 18 (4), pp. 243-254.
Saracevic, Tefko (1999), *Information science*, in Journal of the American Society for Information Science, 50 (12), Pp. 1051-1063.
Schement, Jorge Reina; Lievrouw, Leah (1987b), *A third vision: capitalism and the industrial origins of the information society*, in Schement & Lievrouw (1987a), *Competing visions, complex realities: social aspects of the information society*, Ablex Publishing Corporation, Norwood, New Jersey, pp. 33-45.
Warner, Julian (2000), *What should we understand by information technology (and some hints at other issues)?* in Aslib Proceedings, 52 (9), pp. 350-370.
Winston, Brian (1998), *Media Technology and Society. A History: from the telegraph to the internet*, Routledge.

Creating an XML vocabulary for encoding lute music

Frans Wiering Tim Crawford** & David Lewis***

We describe the development of an XML representation, called TabXML, for encoding historical sources of lute music. These sources employ a special notation type, tablature, that is very hard to understand for non-lutenists. This paper discusses several issues in creating TabXML:

1. what to represent: the *notational meaning* or the *text* of the tablature, and how to represent it;
2. an analysis of the required text-critical markup;
3. provisions for transcription to Common Music Notation and for music retrieval.

The research is situated in the general context of digital critical editions of music.

Keywords: lute tablature, musical sources, critical edition, music information retrieval, XML

1 Introduction

This paper is a preliminary report on a project to define an XML markup language for encoding sources of lute music. It has a more generic aim, too, that is, to stimulate the development of a well-argued concept of digital critical editions of music. In our view, such an edition is best realised as an XML application. Others may be of a different conviction, but their solutions will also have to be able to deal with many of the problems described here.

At a generic level, our research is concerned with scores, instantiations of musical works written down in music notation. Digitisation of musical notation was first attempted in in the early nineteen-sixties. Its main purpose, up to quite recently, was typesetting, traditionally for paper publication but increasingly for online distribution in page-oriented layout.

Exploitation of the information content of such encodings has been around quite a while. Scholarly applications include cataloguing, analysis and transcription (translation of one type of notation into another). Music information retrieval applications, which are widely researched at present, extend the use of musical information to the consumer domain.

Despite the obvious analogy of music notation to text there is no musical parallel to the vigorous argument in literary studies about 'the critical edition in the digital age' [2], and only a few researchers have used the computer's processing power to support critical editing of music. A generation back, Thomas Hall [7] experimented with stemmatics on the Masses of Josquin Desprez (c. 1440-1521); related experiments have been done comparing scores as graphic images [1]. Theodor Dumitrescu [5] has proposed a *Corpus Mensurablis Musice 'Electronicum'*,[1] featuring different views generated from an underlying XML representation of the musical source. Two recent projects may serve to illustrate the required functionality for digital critical editions of music:

– The Online Chopin Variorum Edition[2] (OCVE), de-

*Department of Information and Computing Sciences, Utrecht University, Netherlands. frans.wiering@cs.uu.nl
**Department of Computing, Goldsmiths, University of London, UK. {t.crawford, d.lewis}@gold.ac.uk

vised as a tool for comparing early editions and manuscript sources of the composer's output, needs mechanisms for controlling and visualising variants.
– The Electronic Corpus of Lute Music[3] (ECOLM), aiming at storing and making accessible to scholars, players and others, full-text encodings of lute tablature sources, needs mechanisms for recording and displaying scribal intervention, editorial emendation and normalisation in lute tablature sources, and for making the materials available to non-specialists.

The research reported here is closely tied to the needs of the ECOLM project.

Of the 'open' – freely-available and fully-documented – music notation formats, several are able to handle a certain amount of text-critical information.[4] Two of them are XML applications, MusicXML and MEI. MusicXML's[5] aim is to allow interchange of musical data between notation, analysis, performance and retrieval applications [6]. It supports notational features that are commonly used in critical editions of music, such as editorial accidentals and extra staves for alternative readings.

Because of its interchange purpose, MusicXML concentrates on capturing the *meaning* of the notation; the *textual* aspect of music notation is considered secondary to the meaning. Consequently, the encoding can be strongly hierarchical, and certain information can be inferred. For example, accidentals preceding and dots following a note are descendants of the `<note>` element. Notes in turn are descendants of the `<measure>` element; usually the bar line at the end of the measure is not encoded but inferred from the existence of the `<measure>` element.

This makes MusicXML less suitable for source encoding. The focus in this task is on the written text, which may contain ambiguities and errors that make its meaning less than obvious. In early notation, for example, dots may belong to one note, or separate groups of notes; bar- lines may be misplaced, or be placed following logic other than ours but still contain important information. It is therefore dangerous to impose a strong hierarchy on the written text, as this assumes that the meaning of the notation is completely known. Moreover, such a hierarchy makes it hard to add a layer of text-critical markup that crosses element borders. The extreme solution is to use a representation that only binds together the properties of a single musical character (shape and position); the downside is that translation to a presentation format is much harder.

The purposes of the TEI-inspired Music Encoding Initiative[6] (MEI) are providing a platform-independent format for musical content and metadata, and interchange. While it differs in numerous ways from MusicXML, MEI has the same focus on notational meaning and therefore a similar hierarchical structure. MEI possesses one construction for text-critical markup: several readings of a passage (marked up as `<rdg>`) can be joined in one apparatus entry (marked up as `<app>`). This markup can only be applied to measures and larger units.

1.1 Aim and organisation

The specific problem we address in this paper arises in the context of ECOLM, and concerns the incorporation of text-critical information in the encoding of lute tablature. The solution we are developing is called TabXML. To fulfill its purpose, TabXML must meet the following requirements:

– it encodes the musical *text* of the source;
– it is compatible with the current encoding scheme, TabCode, described in section 3;
– editorial interpretations can be added;
– different views of the source (*e.g.* first scribe, final state, critical edition) can be generated from TabXML;
– (semi)automatic extraction of notational meaning is possible, allowing transcription, music retrieval and interchange with other music encoding languages.

Additional requirements are:

– existing solutions are reused as much as possible;
– in particular, the text-critical markup must be as similar as possible to TEI, to facilitate the encoding of sources that contain both music and text;
– the text-critical markup can be generalised for other types of music notation and (partly) be integrated into other XML-based music encoding languages.

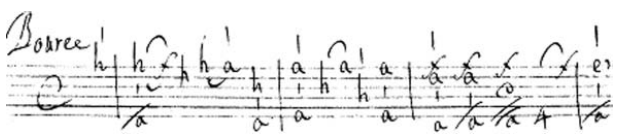
This paper is organised as follows. First, lute music and its notation are described, followed by a description of TabCode, the current encoding scheme. Next follows the main contribution of this paper, a study of text-critical features that occur in lute music, and an investigation of the suitability of TEI markup[7] for encoding these. Then we discuss the processing of the musical content and its requirements, and finally we mention some conclusions and future work.

2 The lute and its notations

The Western European lute is an instrument of great significance in music history. From the end of the fifteenth century until the latter half of the eighteenth century, it 'remained one of the most widely used domestic solo instruments', with an extant repertory of nearly 60,000 pieces [8]. Yet the lute and its music play a comparatively minor part in current musicology: the music is not generally well known and its historical role rarely discussed. One reason for this is the notation that it uses, lute tablature.

Tablature was widely used between about 1450 and 1800, principally for instruments of the lute family, but also for keyboard and other instruments. Tablature generally indicates, by graphical symbols, the finger-placements necessary to perform the music, rather than the 'musical' events that are recorded in common music notation (CMN). Since tablature notations are physical, instrument-specific instructions, they may be convenient for players, but very hard for others to read.

We concern ourselves here with one form of lute tablature, known as French tablature (an example is shown in Figure 1). All lute tablatures indicate which course (i.e. string or pair of strings) to strike with the right hand and which fret on the instrument's fingerboard to stop with left. In French tablature six 'staff-lines' represent the courses, and letters of the alphabet are then placed on a line to indicate the fret to use, *a* being an open string, *b* the first fret, and so on. The time separating one set of notes from the next is indicated using rhythm signs, based on those of CMN, placed above the notes. Since only one rhythm sign is possible at any time, explicit polyphony cannot be directly notated, nor can the duration of a note that is to be sustained while the next note is played. Repeated rhythm signs are usually suppressed.



```
M(C) Qh1 | Eh1Xa/ f1 h2 h1 Qa1 h3Xa | Qa1a6 Eh3 a1
Qh4 a1a6 |
Qf1a2Xa f1a2Xa/ f1d3Xa// EX4 f1 | Qe1Xa/
```

Figure 1. Example of lute tablature [10, fol. 3v], with CMN transcription and TabCode rendition

These core features of the notation might then be supplemented with, for example, signs indicating fingering and ornamentation placed close to the notes to which they apply. As the lute developed over time, it gained extra bass courses in addition to the six played on the fingerboard, taking the total number to anywhere between eight and fourteen. These bass courses are indicated by the placing of additional symbols beneath the staff, usually an *a* with optional slashes or a number.

3 Encoding tablature

ECOLM employs a simple flat-text format called TabCode to encode the tablatures [3]. Since lute tablature is essentially sequential in nature, with one chord following another without notated overlap, the transformation to a stream of text is fairly straightforward and literal.

TabCode, a sample of which is given in Figure 1, consists of a series of 'tabwords' separated by white space. Each word begins with a character indicating the rhythm sign (the initial of the name of that sign: *Q* for quarter note, *E* for eighth, etc.), if present, followed by a letter-number pair for each symbol in the chord, with the number signifying the string and the letter the fret, as in the tablature. An *X* in a tabword separates bass courses from the others. Barlines are indicated using the pipe character '|', and mensural signs, fingerings, ornaments and slurs have their own symbols – for clarity, these are not encoded or transcribed in this paper's examples.

Two approaches to TabXML are possible and offer different advantages. One is to consider TabCode as the flat text of the document, to which the text-critical markup is added. This is the 'simple TabXML' used for the experimental encodings in section 4. The other is to translate TabCode itself to an XML format and then add the text-critical markup layer. This approach is mainly useful when the content of the notation is processed, for example for transcription to CMN or for music retrieval. 'Full TabXML' is briefly described in section 5.

4 Text-critical features and markup

The source we studied for this research is the so-called London Manuscript[8] [10] of the works of Silvius Leopold Weiss (1686-1750), one of the most important composers for the lute, and also one of the last. The manuscript is partly in Weiss's hand; sections written by others contain corrections by Weiss.

We scanned the manuscript for features that require text-critical markup. Subsequently, these were divided into three categories:

1. problems in the source text: errors, missing or illegible information;
2. variant readings: scribal corrections, improvements, or explications;
3. changes to conform to modern usage.

To investigate the suitability of TEI markup, we created a TEI (or TEI-informed) encoding for selected features. Note that the following is a non-comprehensive listing, and that features and markup have application in all music notation encoding, not just in tablature.

4.1. Problems in the source text

Most sources are imperfect, because of errors or physical damage. There are usually clear counterparts for these imperfections in verbal text, except that not only the symbols, but also their precise location is part of the text. We distinguish the following subcategories of problem in the source text.

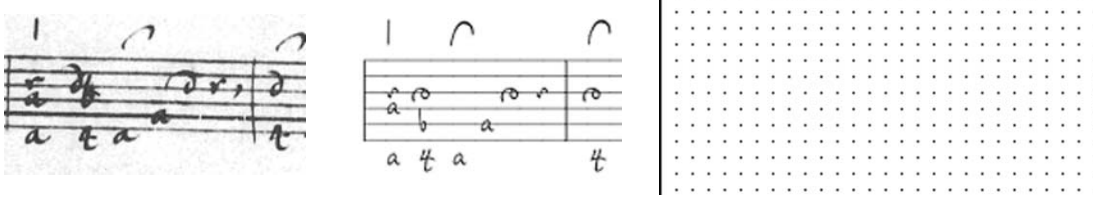
1. Errors. Figure 2 illustrates a very common type of error that is a consequence of the two-dimensional nature of music notation: the symbol is right but the placement is wrong. The error is easy to identify, as the uncorrected passage makes no musical sense. Another common error type is the occurrence of a wrong symbol (letters, rhythm signs, or other).

2. Missing information. Missing duration symbols are often easily noticed, as they usually cause a syntactical error: i.e. wrong measure length. Missing pitches generally do not cause such an error: detecting them is a matter of expertise and there is often more than one solution.

3. Unclear or illegible information. Example 3 shows a blot where a letter is expected. Assuming that the letter is underneath and given the size of the blot, candidates are *a*, *c*, and *e*. Of these candidates, only *a* makes musical sense, as shown in the transcription.

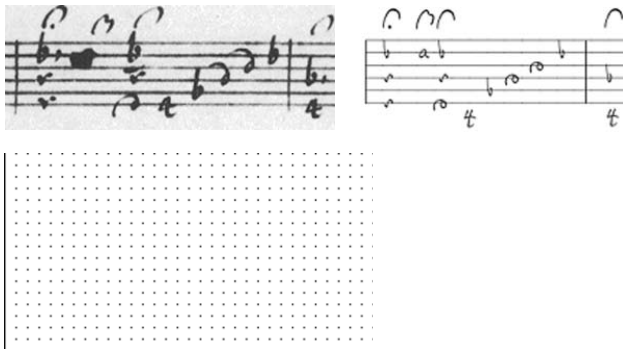
4. Simple corrections. Often mistakes in the source were immediately corrected by the scribe. One such example is shown in Figure 4. In the first chord, the *a* below the bottom line was a mistake that was cancelled by partly overwriting it with an *a* on the line. These corrections belong to a single layer of the text and are therefore very different from the ones described in section 4.2.

5. Damage to the medium. This may vary from local damage to the page to that affecting several leaves or quires. In the London manuscript of Weiss's works, for example, a whole leaf from within an autograph sonata (presumably itself copied by Weiss) was removed; fortunately, it was replaced by one containing the missing music in the hand of the other main copyist [4].



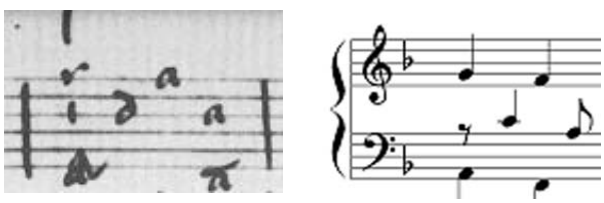
TabXML (1): `Qc3a4Xa d3<choice><sic>b4</sic><corr>b5</corr></choice>X4`

Figure 2. Pitch error (tablature b in second chord is on the wrong line) [10, fol. 31r], with corrected tablature, CMN transcription and TabXML encoding



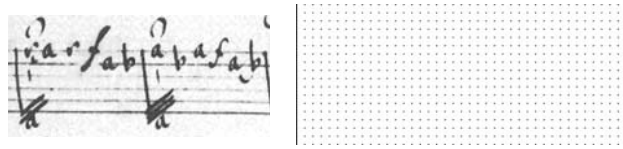
TabXML (2): `E.b2c4c6 S<supplied>a</supplied>2 Eb2c4d6`
TabXML (3): `E.b2c4c6 S<supplied reason='blot'>a</supplied>2 Eb2c4d6`

Figure 3. Illegible symbol in second tabword [10, fol. 13r], with corrected tablature, CMN transcription and TabXML encoding



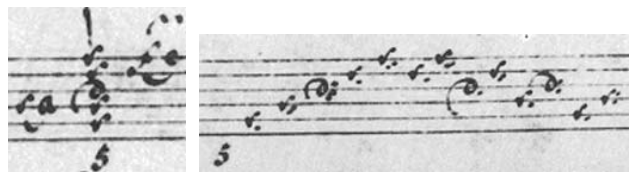
TabXML (4): `Qc1<add>a6</add><del>a7</del> d3`
TabXML (5): `Qc1<subst><add>a6</add><del rend='partly overwritten, not cancelled'>a7</del></subst> d3`

Figure 4. Simple scribal correction [10, fol. 60v], with CMN transcription and TabXML encoding



TabXML (6): `Ec1Xa// a1 c1 <add hand='weiss'>f1</add> a2 b2`
TabXML (7): `Ec1Xa// a1 c1 <app><rdg><add hand='weiss'>f1</add></rdg> <rdg><del hand='other'>f2</del></rdg></app> a2 b2`

Figure 5. Correction in Weiss's hand [10, fol. 19r], with CMN transcription and TabXML encoding



a *b*

TabXML (8): `<app><rdg varSeq='1'>Qc 1c2d3c4c5X5</rdg><rdg varSeq='2' cause='explication'>X5 c5 c4 d3 c2 c1 c2 c1 d3 c2 c4 d3 c5 c4</rdg></app>`

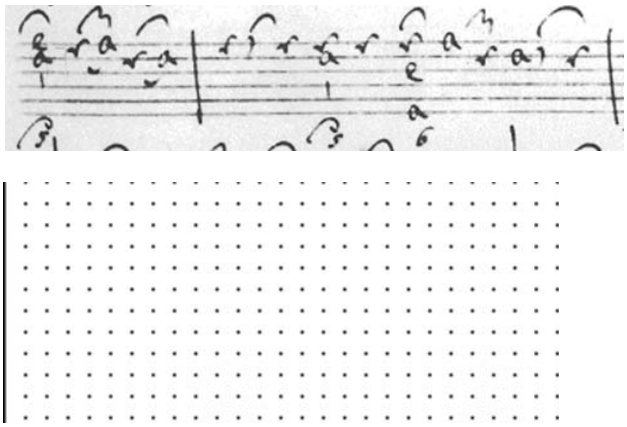
Figure 6. Explication [10, fol. 56v], showing arpeggiation and fingering of the chord

Encodings

Experimental XML encodings are shown below each of the tablature examples. Note that we prefer to mark up the combination of symbol and position as these together constitute the smallest unit of notational information. There is only one deviation from the proposed TEI P5 markup: encoding 5 wraps the `<add>` and `<del>` elements in `<subst>`, indicating the substitution relationship between the two.[9] Generally, no attributes are shown: when this is desirable, these can be used for more precise description of the situation, as shown in encoding 3. For type 4, the `<corr>` element can be used. For type 5, elements `<damage>`, `<supplied>` and `<unclear>` are suitable.

4.2 Variants

Variants to the text of a musical work might be found in various places: as another layer in the same text, as fragments located elsewhere in the same source, or in a different physical source. Large-scale works are an intermediate case as these are often preserved as a full score and separate parts. Even when these clearly belong together interesting variants may occur between them (J.S. Bach sources are famous for this). We can



TabXML (9): `Ee1a2<app><rdg hand='early'>d7</rdg> <rdg hand='late'>X5</rdg></app>`

Figure 7. Adaptation for 13-course lute, original characters preserved [10, fol. 1v], with CMN transcription (notes added later are in parentheses in the transcription)

also make a functional distinction between variants.

1. Correction. Figure 5 shows an inline correction in the first measure, clearly in a different hand – compare the two *f*s. Even though the original note was erased, the difference with the situation described in section 4.1 is that here there are potentially 'two different ways of understanding the text',[10] whereas above there is only one. Larger-scale corrections in the manuscript include cancelled measures, for which replacements are written on another page, and multi-measure insertions.

2. Explication. Music notation is often schematic, and most early music was supposed to be somehow elaborated in performance. Ornaments would be added in addition to those already given in the source, rhythms could be altered and chords arpeggiated. What unites these is that they would never be done twice in exactly the same way. Occasionally, composers would give example realisations, as a suggestion of a possible performance. Figure 6 shows such an explication, which appears elsewhere on the page and includes fingering.

3. Alternatives. Alternative versions may be given for a variety of reasons. One type of alternative is the so-called *ossia*, an extra staff providing a simplified reading for a technically challenging passage, which is often encountered in piano music. In the London Manuscript (fol. 56v), there is the curious case of a incomplete alternative ending to a prelude, possibly showing how it might modulate to another tonality.

4. Adaptation. In adaptation, the musical content is revised. This may happen for various reasons. For example, up to 1717, Weiss used an eleven-course instrument. In that year he acquired a thirteen-course instrument. He adapted some of his works to the new instrument. Figure 7 shows this was done, by replacing some characters by *5* or *6* for the new courses. The older symbols were retained in these measures, in other places they were deleted.

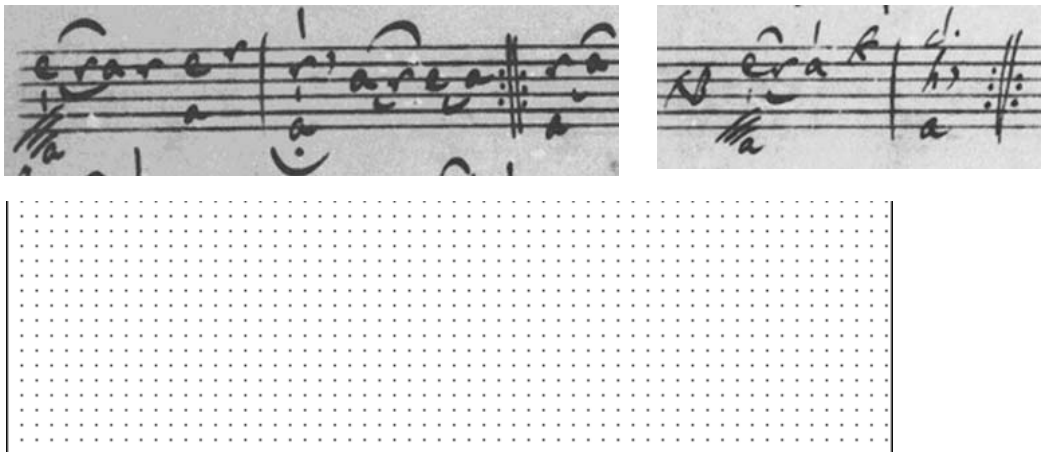
Another option is actual re-composition. Figure 8 shows an example of this: it is the end of the first section of a minuet, which is to be repeated, skipping the last four notes of the second measure. Later, Weiss wrote a new ending for the repeat at the bottom of the page, not only leaving out the notes that lead back to the beginning, but altering the musical content of these measures.

Re-composition can be carried much further than local revision. Up to at least the end of the eighteenth century, works were habitually adapted to suit circumstances, resources, and qualities of the performers. In relatively few cases in early music can we actually be confident that a composition ever reached its final state (as opposed to its last surviving state), which makes access to compositions in all their surviving versions essential for studying and understanding them.

Encodings

Experimental XML encodings are shown below each of the tablature examples. Encoding 6 is very similar to a scribal addition, except the use of the `hand` attribute. Encoding 7 renders the hypothetical situation that the error was an *f* on course 2 – and legible. This markup shows that two successive versions of the text have existed. The markup is also distinctive from encodings 8 and 9, where the later variant does not necessarily cancel the earlier one. In fact, all types in this category may require use of `<app>` and `<rdg>` elements. This markup thus covers a variety of meanings, which then must be differentiated by suitable attributes. On the other hand, the types may be hard to distinguish in practice.

We have not covered the encoding of variants from different physical sources. Because of the generally unfixed state of early musical works, differences may be considerable, including major 'structural' changes. We



TabXML (10):

```
<app><rdg hand='early'>Ee2Xa/// c2 a2 c2 e2a5 c1 | QFc5</rdg>
<rdg hand='late'>Ee2Xa/// c2 Qa2 k1 | H.h3a6</rdg></app>
```

Figure 8. Re-composition [10, fol. 10r], with CMN transcription of later version

```
<choice>
 <orig>QFc2a6 Ea3 c3 e3 a3:||:</orig>
 <reg>(T+\1) Qc2a6 Ea3 c3 e3 a3:|| (T+\2) Fc5a6 ||:</reg>
<choice>
```

Figure 9: Normalisation of notation, encoding the left (older) layer of the tablature shown in Figure 8. Note fol. 10r, encoding of older version, with normalisation of notation.

suspect that this will make the encoding process quite hard and probably not worth the effort. A better strategy seems to be to encode such versions separately and to develop an adequate linking mechanism.

4.3 Changes to conform to modern usage

Given a source that is completely legible and error-free, it may still need adaptations in order to conform modern usage or indeed usability. Tablature itself is a dramatic example of this: even most professional musicians would need a translation into standard music notation in order to understand it. We distinguish the following sub-categories.

1. Adaptation to modern notational conventions. Examples include normalisation of barlines, and writing out a first and second repeat (see Figure 8). Given a good source-text and suitable markup, adaptation is a many-to-one process, that (in principle) can be done algorithmically.

2. Realisation of abbreviations, notably ornament-signs. In printed scores, ornaments are usually left as they are, providing an explanatory table when necessary. For automatic analysis and retrieval thay may need to be expanded.

3. Inference: supplying information about the musical content that is – at best – implicit. This is a one-to-many or a many-to-many process, where knowing the context is vital for assessing the possible alternatives. It may be desirable to specify alternatives, or to recommend one. Examples pertinent to lute tablature are pitch spelling, note duration and voice leading. The most notorious example is the addition of *musica ficta* to Medieval and Renaissance music, where the task is to provide additional sharps and flats to the score that used to be added on the fly, in performance. Inference is necessary for any form of musical content processing, translation to CMN in particular.

Encodings

One attempt at adaptation is shown in Figure 9, using standard TEI P5 tagging (`<orig>` and `<reg>`, wrapped in a `<choice>` element). The TabCode fragments `(T+\1)` and `(T+\2)` indicate the first and second repeat signs. For clarity, we have chosen to encode the entire normalisation, involving adaptation of durations and barlines, and insertion of a tabword, as one editorial intervention, rather than a series of smaller ones. For realisation, the `<abbr>` and `<expan>` elements seem best suited. A possible solution for inference is discussed in section 5.

5 Transcription

The aim of ECOLM is to provide access to lute tablature, for non-lutenists in particular, which means that tablature must be transcribed into CMN and/or made audible. A full version of TabXML should therefore be able to record transcription information, so that this can be created once. This requires breaking up the tabwords into XML elements to which additional data can be added.

Our first goal was a proof-of-concept translation of TabCode into MusicXML. First, a Perl script translates tabcode into full TabXML phase 1. Another Perl script parses the TabXML into a DOM tree, interprets the duration and determines the pitch from the course/fret combination. The output is stored as TabXML phase 2. An XSLT sheet then converts the Tabcode into MusicXML. Finally MusicXML is imported into the Finale music printing program[11] to create a simple CMN version. Experiments have shown that it is feasible to add text-critical markup to this code; another XSLT sheet was used to create a particular view of the tablature before converting it into MusicXML.[12]

Conclusions and future work

Not all of the requirements from section 1.1 have been realised as yet. Bearing in mind that the above survey of text-critical features is only tentative, it seems that the TEI tag-sets for transcription of primary sources and for critical apparatus can be used for marking up for musical sources with only few changes.

An important problem in tablature is the inference of musical meaning needed for different forms of access. Its derivation is not the subject of this paper – but if we want to preserve inferences for later use, it is necessary to extend Tabcode to a full XML representation. This representation must also include metadata and basic rules for interpreting the tablature. We have demonstrated that our current representation is complete enough for primitive transcription into CMN.

Future work includes analysis of other sources, testing, and specification of document model(s) in DTD and/or RELAX-NG. Manually marking up tablatures is clearly not efficient. A dedicated editor for Tabcode has already been created by Christophe Rhodes [9]; support for text-critical markup will be added to it.

Finally, experiments are foreseen with other types of early music notation in collaboration with the MEI project. The ultimate result may be that musical sources can be treated as another document type in TEI, with its own base tag set.

Acknowledgements

We thank Lou Burnard, James Cummings and Perry Roland for their comments during earlier stages of this research. Full TabXML was developed by Frans Wiering during a Visiting Scholarship at CCARH, Stanford University, in Spring 2003. His TabXML research is now supported by EPSRC grant no. GR/T19308/01.

References

[1] Brett, P., & Smith, J. Computer Collation of Divergent Early Prints in the Byrd Edition. *Computing in Musicology* 12 (2001), 251-260

[2] Breure, L., Boonstra, O., & Doorn, P.K. *Past, present and future of historical information science.* Amsterdam: NIWI-KNAW, 2004

[3] Crawford, T. Applications Involving Tablatures: *TabCode* for Lute Repertories. *Computing in Musicology* 7 (1991), 57-59

[4] Crawford, T. Sylvius Leopold Weiss and the London and Dresden manuscripts of his music. Lute Society of America Journal, forthcoming

[5] Dumitrescu, T. *Corpus Mensurablis Musice 'Electronicum'*: Toward a Flexible Electronic Representation of Music in Mensural Notation. *Computing in Musicology* 12 (2001), 3-18

[6] Good, M. MusicXML for Notation and Analysis. *Computing in Musicology* 12 (2001), 113-124

[7] Hall, T. Some Computer Aids for the Preparation of Critical Editions of Renaissance Music. *Tijdschrift van de Vereniging voor Nederlandse Muziekgeschiedenis* 25 (1975) 38-53

[8] Ness, A.J., & Kolczynski, C.A. Sources of lute music. In Sadie, S., Tyrrell, J., eds.: *The New Grove Dictionary of Music and Musicians*, vol 23. Macmillan, London (2001) 3963

[9] Rhodes, C., & Lewis, D. An editor for lute tablature. Paper submitted to Computer Music Modelling and Retrieval Conference, Pisa 2005
[10] Weiss, S.L. *Complete Works for Lute*, vol. 1. Frankfurt, New York, London: Peters, 1983

Notes

1 http://www.cmme.org
2 http://www.kcl.ac.uk/humanities/cch/ocve/final/content/index.html
3 http://www.ecolm.org
4 We use 'text-critical' as a conventient label for features described in the TEI Guidelines under 'transcription of primary sources' and 'critical apparatus'
5 http://www.recordare.com (examined: version 1.0)
6 http://www.lib.virginia.edu/digital/resndev/mei (examined: version 1.5b)
7 We consulted the TEI P5 Guidelines (version 0.1.2), http://www.tei-c.org/P5/Guidelines/
8 British Library Add. 30387
9 Solution taken from http://www.hum.ku.dk/ami/handbook/chapter5.html
10 Daniel Paul O'Donnell, http://sourceforge.net/mailarchive/message.php?msg_id=10186779
11 http://www.finalemusic.com
12 Example output can be found at http://www.cs.uu.nl/people/fransw/tabxml/

Writing history in the virtual knowledge studio for the humanities and social sciences[1]

*Paul Wouters**

This paper presents the goals and theoretical underpinning of a new programme in e-humanities and e-social science in the Netherlands. Recent transformations in communication and information exchange have created new opportunities for researchers in the humanities and social sciences. It is not self-evident, however, in what ways scholars can best use these possibilities while maintaining and further developing their specific roles in academia and society. This is the rationale of a new research programme in the Netherlands, *The Virtual Knowledge Studio for the Humanities and Social Sciences*, hosted by the Royal Netherlands Academy of Arts and Sciences. It aims to support researchers in the social sciences and humanities in the creation of new scholarly practices, termed e-research, as well as in their reflection on e-research in relation to the development of their fields. A core feature of the *Virtual Knowledge Studio* is the integration of design and analysis in a close cooperation between social scientists, humanities researchers, information technology experts and information scientists. This integrated approach should provide insight in the way e-research can contribute to new research questions and methods in the humanities and social sciences.

Introduction

Recent transformations in communication and information exchange have created new opportunities for researchers in the humanities and social sciences. It is not self-evident, however, in what ways scholars can best use these possibilities while maintaining and further developing their specific roles in academia and society. This is the rationale of a new research programme in the Netherlands, *The Virtual Knowledge Studio for the Humanities and Social Sciences*, hosted by the Royal Netherlands Academy of Arts and Sciences. It aims to support researchers in the social sciences and humanities in the creation of new scholarly practices, termed here e-research, as well as in their reflection on e-research in relation to the development of their fields. It is both a novel multidisciplinary re-

1 This work has been developed in close collaboration with Anne Beaulieu, Jenny Fry, Iina Hellsten, Matt Ratto, Andrea Scharnhorst, and Katie Vann.

**Networked Research and Digital Information (Nerdi), NIWI-KNAW, The Royal Netherlands Academy of Arts and Sciences, PO Box 95110, Amsterdam, Netherlands.* Email: paul.wouters@niwi.knaw.nl

search programme with intellectual merits of its own, and a new intellectual and technological infrastructure for the communities of researchers and scholars in the humanities and social sciences.

e-Science is generally defined as the combination of three different developments: the sharing of computational resources, distributed access to massive datasets, and the use of digital platforms for collaboration and communication (Hey & Trefethen, 2002; Nentwich, 2003). The core idea of the e-science movement (most of it still promise rather than practice) is that knowledge production will be enhanced by the combination of pooled human expertise, data and sources, and computational and visualisation tools. In this proposal we use the notion of e-*research* rather than e-*science* to indicate that it is not a matter of importing e-science ways of working into the social sciences and humanities. The humanities and social sciences will develop their own specific ways of integrating the use of networked information and communication technologies (Bijker & Peperkamp, 2002; Bijker et al., 2003; Boonstra, Breure, & Doorn, 2004; Kircz, 2004). This does not have to mean that the difference with natural sciences will become less important. Hence, the generic term e-research is preferable over the notion of e-science.

The humanities and social sciences are no backwater with respect to e-research. For example, archaeology has developed e-science ways of working in its combination of natural science and humanities expertise, its use of sophisticated Geographical Information Systems (GIS) software packages, and its use of expert systems in parts of its research and training. In the field of linguistics, both corpus-based and experimental approaches have led to a transformation of the study of language and the creation of sophisticated research infrastructures. The cognitive sciences are an example of the confluence of natural sciences, social sciences and humanities which drives them into a new experimental direction that relies heavily on computer-based imaging techniques. Economists are interested in modelling and simulation and develop fields like neuro-economics. In sociology, computational research seems to catch on again in the form of new research programmes aimed at, among others, micro-simulations of households and agent-based modelling (Ahrweiler & Gilbert, 1998). Moreover, computerised social network analysis is a well-established tradition in sociology (Wasserman & Faust, 1994). Even in the more traditional fields, many researchers in the humanities and social sciences are adept users of the most advanced tools they can get, as long as the learning curve is not perceived as too steep.

Yet, the implications of e-research for the humanities and social sciences are far from clear. A systematic and critical interrogation of the potential of e-research paradigms and methodologies for the humanities and social sciences has been hampered by disciplinary boundaries between fields, by a relative lack of resources and research infrastructures, and by the dominance of particular computational approaches in the world of e-science. The Studio will address these problems by:

– demonstrating and exploring the potential of additional, non-computational as well as computational, ways of doing e-research
– making disciplinary boundaries more permeable for new scholarly practices
– pooling resources that are available to the scholarly communities in the Netherlands and abroad.

Programme mission

The *Virtual Knowledge Studio* has the following goals:

– to contribute to the design and conceptualisation of novel scholarly practices in the humanities and social sciences
– to support scholars in their experimental play with new ways of doing research and emerging forms of collaboration and communication
– to facilitate the travel of new methods, practices, resources and techniques across different disciplines
– to contribute to a better understanding of the dynamics of knowledge creation.

These goals are intimately connected, one cannot be reached without the other.

A core feature of the Studio is the integration of design and analysis in a close cooperation between social scientists, humanities researchers, information technology experts and information scientists. The programme is integrated at all levels: its goal and mission, its research projects, its internal peer review, its funding acquisition, its collaboration and publication policies and its data management. This does not mean that disciplinary differences will become invisible. On the contrary, we rather expect that these distinctions will be productive and contribute to that creative tension which is the hallmark of innovative research and scholarship. This integrated approach should thereby

provide insight in the way e-research can contribute to new research questions and methods in the humanities and social sciences.

The humanities and social sciences are not a unified set of knowledge practices. Moreover, methodologies and techniques do not travel easily from one field to another. This has a direct bearing upon the development of e-research tools, practices, infrastructures and institutions. The equipment of many academic scholars with the tools and 'play space' they need to independently assess the merits of e-research has moreover been hindered. Lack of funding has decreased the space for advanced instrumentation and support staff. Moreover, ICT has been standardised within the paradigm of office automation and therefore lacks many features that would be useful for scholarly work. In so far as ICT has been tailored to research needs, it has been based on computational research and often assumes mathematical and programming skills on the part of the researcher. In many fields scholars have different needs, such as the representation of ill-defined data, analysis-oriented visualisation of manuscripts and multimedia sources, and specific source-oriented analytical tools. These needs are often not met by standard computational and mathematical analytic methods. This is especially true for many fields in history and archeology.

The meeting of e-research and the academic scholar is moreover problematic because it is far from clear whether the present needs of the scholar can be met by e-research at all. Important fields in the humanities and social sciences are characterised by a huge epistemic diversity; by specific, sometimes person-bound, roles of the researcher; by the lack of consensus about the research agenda in a host of specialties; by a relatively low-tech research environment (often aggravated by the scarcity of university funding); by the specificity of writing and reading as features of knowledge creation; and by a historically grounded and relatively large share of solitary research practices (Becher 1989; Whitley 2000). In all these dimensions, many fields seem ill-suited to become enthusiastic adopters of the e-science paradigm as it now stands. If e-research should make sense to a variety of specialties in the humanities and social sciences, new non-computational and computational paradigms of e-research need to be developed.

The Studio will therefore orient itself to a critical interrogation of the very notion of e-research, by taking seriously the intellectual and social characteristics of the humanities and social sciences and the implications of these characteristics for the hermeneutics of e-research as a prospective intellectual and technical horizon. At the same time, existing knowledge practices in the humanities and social sciences should not be taken for granted. There is ample space for enhancement indeed (see for example the NWO programs focused on new research practices in the humanities and social sciences (NWO, 2000, 2001a, 2001b)). This must in the end be enacted by the research communities themselves, which is why we wish to engage them in the Studio. Advanced research projects and the discussion of these projects in seminars and Summer Schools can play a catalytic role in this development by the involvement of new generations of researchers and students (PhD and Research Masters). Each research project will result in contributions to the pool of research resources in the form of scientific and technical publications, research methodologies and techniques, software tools, organisational protocols, or best practice manuals, and freely downloadable data and tools. Because the research projects in the Studio will not be developed *for* but *with* scholars in the humanities and social sciences, we expect that the lessons learned in this research programme will have a lasting effect on academia in the Netherlands.

This may be reinforced by the host of recent new initiatives in the areas of digitisation, Web based repositories and archives, digital libraries, and collaboratories, in the Netherlands as well as abroad. New programmes such as the Dutch national initiative DARE[2] and the NWO programme CATCH[3] attest to this. Indeed, digitisation of the humanities has been on the agenda for a number of years (NWO, 2000). Recently, a consensus has emerged about the need for a national data archive in the humanities and social sciences (DANS) (see for the social sciences SWR, 2003). This coincides with a rethinking of the social and cultural roles of the humanities in the Netherlands (Bijker & Peperkamp, 2002; NWO, 2002) and abroad (*e.g.*. Ang & Cassity, 2004). The Studio will not duplicate these efforts. Its research will also not try to

2 http://www.surf.nl/themas/index2.php?oid=18

3 http://www.nwo.nl/subsidiewijzer.nsf/pages/NWOP_66EUM7?Opendocument

take over the responsibility of R&D departments of data archives, repositories and academic and research libraries. To the extent that these R&D efforts will take more shape, the Studio will take initiatives to cooperate in joint research projects on common themes. We expect that these projects will focus on the role of data and data standards in scholarly work.

Complementary relationships also will be developed with scholars engaged in methological research in social science and humanities university departments, and with research groups in humanities computing. The Studio will not try to interfere with already well-established methodological traditions and research programmes in Dutch universities. Monodisciplinary methodology is the responsibility of the relevant research groups, not of the Studio. Rather, the Studio will contribute to the exchange of methodologies between different research traditions by creating an experimental space and repository of e-research related methods and tools. In e-research it is sometimes less clear what a research method actually entails than in traditional research contexts. For example, the difference between tools for communication and tools for analysis may become blurred. This is especially true for collaborative analytical and annotation tools, a niche area that may be worthwhile for the Studio researchers to explore. This area is usually not yet covered in more traditional methodological research. The Studio moreover aims to contribute to the methodological development of the study of e-research itself. To this end, the Studio develops a concentrated research effort in three specific methodological domains.

Research themes

The Studio will concentrate its work in three *research themes*:

- Data and Digital Information: the role of data, digital information and data standards in scholarly research
- Networked Research: novel forms of collaboration and communication in the humanities and social sciences
- Virtual Institutions: the emergence and dynamics of new institutional arrangements in e-research.

Data and digital information play complex roles in research in the humanities and social sciences (Arzberger et al., 2004; Boonstra et al., 2004; SWR, 2003). This creates particular challenges for the application of e-research methods and techniques, especially if complex and fuzzy data sets are involved (*e.g.*. visual data, music, complex texts). The increased availability of digital resources, data and collections, partly the result of digitisation of cultural heritage and of administrative databases, affects the very core of humanities and social science research by changing existing research objects and creating new ones. The theme Data and Digital Information will address the question of which characteristics these new research objects will and should have, and how they may reconfigure scholarly research. What type of questions will be foregrounded and which questions may become less central? Which assumptions are built into the new epistemic objects and how may they influence the boundaries between scientific specialties? We will also pay attention to the specificity of qualitative data. They are often more fuzzy and less easy to standardise. The Studio research will strive to complement existing research into scientific and scholarly data and data standards by focusing on the epistemic and social role of data and data sources in the humanities and social sciences. Additional attention will be given to issues of data sharing and the specific problems related to the use of Web data in scholarly work (Wouters & Schröder, 2003).

The humanities and social sciences are a particularly interesting area to study the development of scientific and scholarly collaboration because the variation of forms of collaboration and non-collaboration is so huge (Fry, 2003). Virtually every possible configuration is practiced in one field or another. The Studio research in the theme Networked Research will focus on the way the new media interact with forms of collaboration and communication. It moreover aims to support scholars with building new forms of collaboration (*e.g.*. collaboratories) and communication (*e.g.*. new Web site conceptions). A key issue concerns the ways the dynamics of collaboration are affected by mediation by digital communication networks. How does the technological possibility intersect with traditional human needs for communication? The implications of collaborative work for the resulting knowledge products will also be studied. How are forms of knowledge affected by the way they need to be communicated? Which types of intellectual work seem amenable to virtualisation and digitisation?

In e-research, digital infrastructures and emergent institutions play a crucial role (Bowker & Star, 1999). Collaboratories, research infrastructures or the lack thereof, digital libraries, digital repositories and col-

lections, and new venues for scholarly publication directly influence the extent to which scholars in the humanities and social sciences can effectively make use of new research possibilities. Given the recent emergence of e-research, the consequences of the accompanying institutional rearrangement is not yet well understood. The theme Virtual Institutions will explore which institutional arrangements are conducive to the humanities and social sciences. This should help to understand the specificities of institution building in the humanities and social sciences. Important questions are also: how does the textual nature of digital infrastructures affect textual practices of researchers and scholars? How does infrastructure sustain various levels of formalisation and circulation of knowledge and information? How universities and research institutes have organised their systems of quality control and accountability may have a profound effect on knowledge creation because of its impact on the criteria of scientific and scholarly quality and integrity. What are the implications of new e-research information infrastructures for regimes of quality control in universities and research institutes?

Methodological foci

The Studio will develop methodological innovation of the study of e-research. Of course, this should be relevant to other researchers in the humanities and social sciences as well. The Studio focuses on those methodologies that (1) are not yet well covered by methodologists in social sciences and humanities at the universities, and (2) are particularly relevant for the study of scientific and scholarly knowledge practices. Three *methodological foci* will be given priority in the first three years of the Studio:

– Virtual Ethnography
– Web Archiving for scholarly research
– Simulation in e-research.

Virtual ethnography extends the notions of field and ethnographic observation from the exclusive study of co-present and face to face interactions, to a focus on mediated and distributed interactions (Beaulieu, 2004, forthcoming 2005; Hine, 2000, forthcoming 2005; Howard, 2002; Mason, 1996). The key research question in the Virtual Ethnography focus will be how ethnography can be pursued in mediated settings. This research should establish which aspects of ethnographic research are challenged in particular in the shift from face to face interaction to mediated digitised interaction. This will make clear how ethnography can be conceived as a flexible practice, while remaining recognisable as a specific methodology (Geertz, 1983). More specific questions are which new concepts of 'field' or 'research site' are needed for virtual ethnography, how virtual elements can be integrated in traditional fieldwork, and which new ethical issues arise in the practice of virtual ethnography.

The Web Archiving focus will develop a new methodology for systematic, longitudinal analysis of Web sites that are produced in the sciences and humanities (Foot & Schneider, 2003; Foot & Schneider, 2002). Presently there is no clear way how to make Web data available for scholarly research. Libraries and archives are now only beginning to develop concepts that enable the medium- and long-term archiving of Web sites. The work in this methodological focus will combine in-depth qualitative Web site analysis with large-scale comparative 'surface analysis' of the Web. The central question is which specifications and analytical tools are needed for the extraction of meaningful data sets for research in the humanities and social sciences from the flood of raw Web data. The Studio will not take it on to crawl and archive the Web itself: this must be the responsibility of libraries and archives. It will however in cooperation with the WebArchivist organisation[4], the Internet Preservation Consortium[5], the Internet Archive[6], and the nascent European Internet Archive, develop methods and techniques to conceptualise Web archives in such a way that they can produce datasets for social science and humanities research.

The aim of the research in the methodological focus Simulation is to develop further expertise in simulations and systematic reflection on the heuristic value of modelling and simulation for theory building in the social sciences and humanities (Burenhult, 2002; Gilbert & Troitzsch, 1999; Schweitzer, 2002). The extended use of data visualisation technologies and virtual reality techniques in simulation research methods is

4 http://www.webarchivist.org/
5 The author is member of the scientific committee of the IPPC.
6 http://www.archive.org/

often seen as one of the hallmarks of e-science (Berman, Fox, & Hey, 2003). The respecification of general simulation models for research questions in the social sciences and humanities will be central in this focus. Both agent-based and network-oriented models and simulations will be included (Ratto & Scharnhorst, 2004; Scharnhorst, 1998; Scharnhorst, 2001). The study of the heuristic and epistemic value of modelling and simulating as a research strategy in both the humanities (*e.g.*. language variation processes) and social sciences (*e.g.*. social selection) is intrinsic part of this respecification.

Research methods and organisation

Three modes of enquiry are central in the Studio: thinking, observing and playing. These three metaphors capture the interplay we expect between thorough analysis and more experimental, playful design of new tools and practices. The design of new tools is never only a technical job. Opening up new possibilities for humanities and social science scholars with the help of advanced networked information and communication technologies implicates the rethinking of old research questions, questioning established research methods and techniques, and asks for the intellectual courage to try out new forms of scholarly work. This is why we emphasise that the Studio will not in the first place help design new tools but rather new scholarly practices.

To realise its dual mission of increasing our understanding of e-humanities and e-social science, and of supporting scholars to make use of e-research, the Studio has two interrelated modules: the Analytic Centre (AC) and the Construction Platform (CP). These facilitate long-term research based on a clear intellectual agenda (AC) combined with flexible short-term projects created in response to the changing needs of researchers at universities and research institutes (CP). For this reason, all Studio research projects will have a complex blend of curiosity-driven and application oriented goals (Ang & Cassity, 2004). All projects in the CP result from, and are led by, partnerships with external research groups. Whereas the CP helps create new epistemic objects and practices in the humanities and social sciences, both inside and outside of the Studio, the AC studies this process. To facilitate this, the AC is responsible for the creation and maintenance of the Studio's inhouse knowledge database.

For the scholars who are the client-partners of the Studio, for example historians, the design work must lead to useful insights in the form of concrete deliverables, such as new protocols, best practice manuals, new software tools, perspectives on new analytical techniques to answer old questions, and new research questions in their fields. For the researchers at the Studio, this design work is also a mode of enquiry into the process of knowledge creation. In other words, the Construction Platform is a field laboratory in which different scholarly practices and configurations are tried out and assessed on their consequences. This will, we hope, lead to a better understanding of the characteristics of knowledge creation as a cultural and social process. The researchers in the Analytic Centre have a special responsibility to link up the results of the CP to the scientific and scholarly literature in the fields of information science, science & technology studies, and communication sciences. To facilitate the management of this type of research, the AC reviews the research in the Studio on its contribution to basic knowledge about the process of knowledge creation. The CP will specifically examine the utility of the Studio research for scholars in the humanities and social sciences based in universities and research institutes in the Netherlands.

Observation, with all the advanced observation tools available in social and cultural analysis, is central to the Studio. This may involve the participant observation of prototypes of new infrastructures (such as collaboratories or Grid computing for social science), but may also entail the systematic observation of mundane processes in research in the social sciences and humanities. This is important to counteract the danger of bias in favour of 'the new new thing' (Lewis, 2000; Woolgar, 2002). We can only put the promise and practice of e-research into perspective by taking distance from the claims and critically interrogate both the promise and the practice (cf. Wouters & Schröder, 2003). This also holds for the innovative projects that are conducted within the Studio itself. Since these are oriented to the exploration of new modes of inquiry, they run the danger of biasing the novel over the traditional. Reflexive self-observation in different forms is therefore an important element of the research cycle in the Studio.

Forms of collaboration

The Studio maintains two forms of long-term collaborative relationships with researchers in the humanities

and social sciences: *partnerships* and *collaboratories*. Partnerships are the main vehicle to intensify cooperation with scholars in the Netherlands around projects. The partnerships are formed on the basis of common research projects and include detailed arrangements about project leadership, project management and dissemination of the results and data. The depth of the partnership may vary. Some partnerships may limit themselves to a particular research project, others may amount to common research programmes, dual career possibilities for postdocs and PhD students, sustained combined acquisition of external funding, and chairs in particular areas of e-research. The Studio collaboratories are aimed at mobilising international expertise in specific areas, especially where there is a lack of experts in the Netherlands. The Studio will start off with four collaboratories. The research in the Methodological Foci will be organised in collaboratories with international research groups that have specialised in the specific methodology: virtual ethnography, Web archiving for scholarly research, and simulation. In addition, the Studio will organise its work on Web data and Webometrics also in the form of a collaboratory.

References

Ahrweiler, P., & Gilbert, N. (1998). *Computer Simulations in Science and Technology Studies*. Berlin, Heidelberg, New York: Springer Verlag.

Ang, I., & Cassity, E. (2004). *Attraction of Strangers. Partnerships in Humanities Research*. Sydney: Australian Academy of the Humanities.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., et al. (2004). SCIENCE AND GOVERNMENT: An International Framework to Promote Access to Data. *Science, 303*(5665), 1777-1778.

Beaulieu, A. (2004). Mediating ethnography: objectivity and the making of ethnographies of the internet. *Social Epistemology, 18*(2-3), 139-164.

Beaulieu, A. (forthcoming 2005). Sociable Hyperlinks: an ethnographic approach to connectivity. In C. Hine (Ed.), *Virtual Methods: Issues in Social Research on the Internet*.

Berman, F., Fox, G., & Hey, T. (2003). The Grid: past, present, future. In F. Berman, G. Fox & T. Hey (Eds.), *Grid Computing. Making the Global Infrastructure a Reality* (pp. 9-50). Chichester, West-Sussex, UK: John Wiley & Sons.

Bijker, W., & Peperkamp, B. (2002). *Geëngageerde geesteswetenschappen. Perspectieven op cultuurveranderingen in een digitaliserend tijdperk* (No. Achtergrondstudie nr. 27). Den Haag: Adviesraad voor het Wetenschaps- en Technologiebeleid.

Bijker, W., Schurer, K., Stronks, E., Uszkoreit, H., Wittenburg, P., & Woolgar, S. (2003). *Building the KNAW International Research Institute on* e-*Science Studies in the Humanities and Social Sciences (IRISS)*. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Boonstra, O., Breure, L., & Doorn, P. (2004). Past, Present and Future of Historical Information Science.

Bowker, G. C., & Star, S. L. (1999). *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: MIT Press.

Burenhult, G. (Ed.). (2002). *Archaeological Informatics: Pushing The Envelope CAA2001*. Oxford: Archeopress.

Foot, K., & Schneider, S. (2003). Alliances or Antagonies? Hyperlinks & Associative Relations in Web Sphere Analysis. *Journal of Computer-Mediated Communication*.

Foot, K. A., & Schneider, S. M. (2002). Online Action in Campaign 2000: An Exploratory Analysis of the U.S. Political Web Sphere. *Journal of Broadcasting and Electronic Media, 46*(2), 222-244.

Fry, J. (2003). *The cultural shaping of scholarly communication within academic specialisms.* University of Brighton, Brighton.

Geertz, C. (1983). *Local Knowledge*. New York: Basic Books.

Gilbert, N., & Troitzsch, K. G. (1999). *Simulation for the Social Scientist*. Buckingham, Philadelphia: Open University Press.

Hey, T., & Trefethen, A. E. (2002). The UK e-Science Core Programme and the Grid. *Future Generation Computer Systems, 18*(8), 1017-1031.

Hine, C. (2000). *Virtual ethnography*. London: Sage.

Hine, C. (Ed.). (forthcoming 2005). *Virtual Methods*: Berg.

Howard, P. (2002). Network ethnography and the hypermedia organization: New organizations, new media, new methods. *New Media & Society, 4*(4), 551-575.

Kircz, J. (2004). e-*based humanities and* e-*humanities on a SURF platform*. Utrecht: Stichting SURF.

Lewis, M. (2000). *The fnew new thing: A Silicon Valley Story*. New York: Norton.

Mason. (1996). Moving Toward Virtual Ethnography. *American Folklore Society Newsletter, 25*(2), 4-6.
Nentwich, M. (2003). *Cyberscience. Research in the Age of the Internet*. Vienna: Austrian Academy of Sciences Press.
NWO. (2000). *Een digitale bibliotheek voor de geesteswetenschappen*. Den Haag: NWO, Gebiedsbestuur Geesteswetenschappen.
NWO. (2001a). *Geesteswetenschappen Strategienota 2002-2005*. Den Haag: NWO.
NWO. (2001b). *Maatschappij- en Gedragswetenschappen 2002-2005*. Den Haag: NWO.
NWO. (2002). *Culturele vernieuwing en de grondslagen van de geesteswetenschappen*. Den Haag: NWO.
Ratto, M., & Scharnhorst, A. (2004). *From numeric to metaphoric simulations – attempts to draw a simulation landscape. Paper given at Epistemological Perspectives on Simulation . A Cross-Disciplinary Workshop. Koblenz, Germany, July 1-2, 2004*.Unpublished manuscript.
Scharnhorst, A. (1998). Citation-Networks, Science Landscapes and Evolutionary Strategies. *Scientometrics, 43*(1), 95-106.
Scharnhorst, A. (2001). Constructing Knowledge Landscapes within the Framework of Geometrically Oriented Evolutionary Theories. In J. Kriz (Ed.), *Integrative Systems Approaches to Natural and Social Sciences Systems Science 2000* (pp. 505-515). Berlin: Springer.
Schweitzer, F. (Ed.). (2002). *Modeling complexity in economic and social systems*. New Jersey: World Scientific.
SWR, S. S. C. (2003). *Networked Data Services: Towards a Future Data Infrastructure for the Social Sciences in the Netherlands*. Amsterdam: KNAW.
Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, NY: Cambridge University Press.
Woolgar, S. (2002). *Virtual Society? Technology, Cyberbole, Reality*. Oxford: Oxford University Press.
Wouters, P., & Schröder, P. (2003). *Promise and Practice in Data Sharing*. Amsterdam: NIWI-KNAW.

Towards a genealogical ontology for the Semantic Web

*Ivo Zandhuis**

Genealogy is an interesting domain for demonstrating the possibilities of the Semantic Web. As a first step towards such a demonstrator this paper introduces an ontology for the genealogical domain. First the existing standards are investigated and a short introduction into the Semantic Web is presented. After that two ontologies are introduced. The first ontology models the capturing of personal information from an (archival) resource. The second ontology models the personal information itself and can be used to capture genealogical information. The most important aspects of the ontologies introduced are (1) the possibility to model the relationship with the source from which the information is obtained, (2) the possibility to model the relationship with the agent responsible for the information and (3) the ion from nomenclature and time-representation in various cultures.

1 Introduction

The World Wide Web is intensively used for genealogical research. People from all over the world investigate their ancestry by searching databases and publish their genealogy on a home-made web site. The information on these platforms however is not structured to facilitate intelligent cross-platform searching.

Family constructions are often used to introduce computing scientists into the field of Artificial Intelligence. The Semantic Web combines the insights in Artificial Intelligence with the possibilities of internet. The automatic construction of genealogies is therefore an obvious demonstrator of the Semantic Web.

This paper is the first step towards this demonstrator discussing a standard for exchanging genealogical data in the form of an *ontology*.

1.1 Goals

A standard for exchanging genealogical data could be used in various situations. The data can be used for finding a person in genealogical research or primary sources published on the Web. With enough data complete genealogies could be derived automatically. When complete reconstructions of populations have been derived, new demographic questions can emerge and be answered.

1.2 Requirements

To reach the goals mentioned above, the standard should have the following properties. Firstly, users should get correct data only. In a system based on the ontology, various integrity checks must be possible, like the check that a person has died after he has been born, or that he can not be over 120 years old. Secondly, the relations between the genealogical data and sources that prove the assertions must be stored, as well as the agent responsible for publishing the assertions. Finally, the ontology must facilitate extensibility for different cultural approaches to persons names and dates and research questions in the future.

*(ivo@zandhuis.nl) Ivo Zandhuis Research & Consultancy

1.3 The existing standards

This paper is not the first initiative in developing a technology for exchanging genealogical data. The most important existing standard in this field is GEDCOM [1]. The latest official version of GEDCOM dates back from 1995 and facilitates migration of data from one system to another.

Another important development in the field of genealogical data is the datamodel developed and published by GENTECH [2]. The datamodel functions as a reference model for programmers of genealogical software. GEDCOM correctly encodes genealogical information. However the technical implementation is outdated, the reference to a source is done with a single string and its oriented on the situation in the US. Like in GEDCOM and GENTECH, the ontology introduced here uses events to model information about birth, death and marriage.

2 The Semantic Web

2.1 Introducing the Semantic Web

This section gives a short introduction into the Semantic Web. A more extensive introduction can be found in [3].The information on the Semantic Web is described in formal languages of different types and levels of expression: Resource Description Framework (RDF)/ RDF Schema [4] and the Web Ontology Language (OWL) [5]. Basis of all these languages is the concept of the *triple*.Take for example the sentence 'Joe has a mother called Sue.' On the Semantic Web this relation is modelled as the triple:

```
('Joe', hasMother, 'Sue')
```

The three parts of a triple are called Subject ('Joe'), Property ('hasMother') and Object ('Sue'). Subjects and Objects are instances of a certain Class. Classes, Properties, Subclasses can be defined in a similar way as in Object Oriented Programming. The declaration of all the Classes, Properties and their relations is called an *ontology*. More precisely: Guarino [6] states that an ontology refers to an engineering artifact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words. The vocabulary is defined in terms of Classes and Properties. For every domain an ontology can be developed; this paper introduces an ontology for the genealogical domain. The ontology and data based on the scheme can be distributed to agents (human or automatic) in XML.

2.2 Why the Semantic Web?

Now why should we use SemanticWeb technology, when GEDCOM is both a correct and widely used standard for exchanging genealogical information?

The Semantic Web languages make extensions possible with concepts needed in all types of future research. Without changing the existing ontologies, researchers can add the extensions themselves. The second reason is that the Semantic Web adds the possibility to use standardized, existing SemanticWeb tools for reasoning about personal information and family relations. Semantic Web tools are built to derive information encoded in an language, i.e. a language not specially developed for genealogical information. Therefore more people are developing tools and better tools are available.

Finally the Semantic Web standards are developed solving known problems in Computing Science. So Computing Scientists discuss the typical problems of formal languages, computability and distribution. The users of the Semantic Web (for instance in the genealogical domain) can use the technology without worrying about these problems, knowing they are dealt with.

3 Modelling information

To reach our goals two ontologies are developed: genont and srcont. The first ontology models the personal information, the second ontology models the source with the personal information it contains. In actual practice three files are stored: one with the definition of the concepts for personal information in the genont-ontology, one with the definition of the concepts for resources in the srcont ontology and a datafile which contains the actual data. The actual data could be either the transcription of a genealogical resource or a digital publication of genealogical research.

3.1 Persons: genont

The genont ontology models personal information of interest for genealogical research. There are two types of persons: female and male. Furthermore a person has relations with names, events (birth, marriage, death), other persons (family-relations) and other attributes (occupations and addresses).

3.2 Sources: srcont

The srcont ontology models the reference to an original source in a repository, such as an archival item or a publication. To archival material is referred by stating the repository, name of the archive, a description of the referred item with a number, a title and a date. A publication is described with author, title, imprint and pages. Besides the reference to the source the containing personal information is captured. This is done by using the constructions of the genont ontology. The genont ontology is therefore imported into the srcont.

3.3 Datafile containing information of a person in a source

For capturing the personal information stored in the resource, a datafile is constructed based on the srcont ontology, extended with the genont ontology.

3.4 Datafile containing genealogical research

A datafile based on the genont ontology can be used to publish the result of genealogical research. In that case only the genont constructions are used. The relations with the sources can be maintained by stating that the persons mentioned in a datafile based on srcont (extended with genont) and those mentioned in the datafile based on genont are equivalent. This can be done with the standard OWL-constructor equivalentClass.

4 The ontologies

In this section the actual ontologies are defined. The OWL Syntax is used [5], which is easier to read then the XML-syntax.

4.1 Responsible agent

There are standard-language constructions for capturing the agent (human, institution, software agent) responsible for publishing the information. This could either be the agent that constructed a genealogy (in a datafile based on genont) or the agent that made the transcriptions of a source available (in a datafile based on srcont).

```
AnnotationProperty(<http://purl.org/
dc/elements/1.1/creator>)
AnnotationProperty(<http://purl.org/
dc/elements/1.1/date>)
AnnotationProperty(<http://purl.org/
dc/elements/1.1/publisher>)
```

4.2 Description of the classes and its relations

4.2.1 Source (srcont)

The srcont ontology defines the concept *item* which has a relation with the description of the resource (repository, name of the archive etc.) and has a relation with the *content* of the resource. This content is stored according to the model defined in the genont ontology.

```
Class(item partial
      restriction(srcref cardinality(1))
ObjectProperty(srcref
      domain(item)
      range(archref))
ObjectProperty(content
      domain(item)
      range(person))
```

The reference to an archival item is made with four components: the repository, the title of the archive, the description of the archival unit and the date of this unit.

```
Class(ref partial)
Class(archref partial ref)
DatatypeProperty(repository
      domain(archref)
      range(xsd:string))
DatatypeProperty(archtitle
      domain(archref)
      range(xsd:string))
DatatypeProperty(desc
      domain(archref)
      range(xsd:string))
DatatypeProperty(unitdate
      domain(archref)
      range(xsd:string))
```

4.2.2 Person (genont)

The Class *person* is the basic-entity in the *genont* ontology. All other entities described below relate to this Class. The following declaration states that a *person* has a name, is born, and by definition has exactly one (biological) father and exactly one (biological) mother. The ontology distinguishes two subclasses of a person: *male* and *female*. Situations where the sex is unknown, the *person*-class itself can be used.

```
Class(person partial
      restriction(hasName cardinality(1))
      restriction(isBorn cardinality(1))
      restriction(hasFather cardinality(1))
      restriction(hasMother cardinality(1)))
Class(male partial person)
Class(female partial person)
DisjointClasses(female male)
```

4.2.3 Name (genont)

The main relation a *person* has, is the relation with a name. Various cultures have different ways of constructing a personsname. For every culture a subclass of the class *name* can be defined in the ontology to support cultural differences. In this paper the construction for Dutch names, mainly before 1800, is elaborated. A typical Dutch name from this era has four parts:

– Voornaam (forename, *e.g.* the English equivalent "Joseph")
– Patroniem (patronymic)
– Tussenvoegsel (*e.g.* "van" of "van der")
– Achternaam (surname)

All of these parts are properties of the class *dutchName*.

```
Class(name partial)
Class(dutchName partial name)
DatatypeProperty(voornaam
      domain(dutchName)
      range(xsd:string))
DatatypeProperty(patroniem
      domain(dutchName)
      range(xsd:string))
DatatypeProperty(tussenvoegsel
      domain(dutchName)
      range(xsd:string))
DatatypeProperty(achternaam
      domain(dutchName)
      range(xsd:string))
```

The class *dutchName* can be related to a person with the following relations; i.e. are properties of the class *person*.

```
ObjectProperty(hasName
      domain(person)
      range(name))
ObjectProperty(hasDutchName
      domain(person)
      range(dutchName))
SubPropertyOf(hasDutchName hasName)
```

4.2.4 Event (genont)

A person goes through various events in his life. Typical events of genealogical interest are birth, death, marriage, baptism, burial or cremation. Here only the birth-event is described. Other events have the same construction.

```
Class(event partial
      restriction(certainty cardinality(1))
      restriction(place cardinality(1)))
Class(birth partial event)
ObjectProperty(isBorn
      domain(person)
      range(birth))
```

An *event* has a spatial and a temporal dimension, both of which can be related to the event. The property that relates the event to a *place* refers to the name of the place. Eventually it should refer to a formal definition of a geographical unit, in a certain period, most likely defined in a specific geographical ontology. Until then the name of the place is a simple string.

```
DatatypeProperty(place
      domain(event)
      range(xsd:string))
```

Time is modelled with a reference to the name of a moment: a date. In various cultures and eras the name of a moment is constructed in another way. In this ontology the Dutch culture of the gregorian date is elaborated.

```
Class(date partial)
Class(gregorianDate partial date
      restriction(day cardinality(1))
      restriction(month cardinality(1))
      restriction(year cardinality(1)))
DatatypeProperty(day
      domain(gregorianDate)
      range(xsd:string))
DatatypeProperty(month
      domain(gregorianDate)
      range(xsd:string))
DatatypeProperty(year
      domain(gregorianDate)
      range(xsd:string))
```

A moment is always a time interval. Maybe small, like the interval between the start and end of a minute, sometimes longer, like the interval between the first of

January and the last of December of a year. The relation between an event and the time interval in which the event takes place are inspired by [7]. Mostly the *event* takes place during the specified time-interval (like 'the birth took place on October 29th, 2003' or 'He died in 1985'), but sometimes the source only reveals that an event took place before or after a specific interval. In the ontology the following three relations are defined: during, before and after.

```
ObjectProperty(certainty
      domain(event)
      range(date))
ObjectProperty(after
      domain(event)
      range(gregorianDate))
ObjectProperty(before
      domain(event)
      range(gregorianDate))
ObjectProperty(during
      domain(event)
      range(gregorianDate))
SubPropertyOf(after certainty)
SubPropertyOf(before certainty)
SubPropertyOf(during certainty)
```

4.2.5 Family-relations (genont)

Obviously in genealogy the family-relations are very important. These relations are modelled as properties of a person. As we have seen before there are two family-relations needed in all situations: the hasFather property and the hasMother property. All other types of family-relations can be constructed, *e.g.* the hasChild property.

```
ObjectProperty(hasParent
      inverseOf(hasChild)
      domain(person)
      range(person))
ObjectProperty(hasChild
      inverseOf(hasParent)
      domain(person)
      range(person))
ObjectProperty(hasFather
      domain(person)
      range(male))
ObjectProperty(hasMother
      domain(person)
      range(female))
SubPropertyOf(hasFather hasParent)
SubPropertyOf(hasMother hasParent)
```

5 Example

At [http://www.zandhuis.nl/sw/genealogy/] an example can be found of datafiles containing genealogical information and its sources. Besides the ontologies in XML-syntax, there are two datafiles: one containing the source-material (source.owl.rdf) and one containing the conclusions of the research (conclusion.owl.rdf). In the first file transcriptions of data from the archive is captured. The last file refers to the sources that prove the assertions made.

Conclusion

The Semantic Web is very suitable for publishing genealogical data in an open and extensible way. In this paper a first attempt is presented for a Genealogical Ontology, that can start the discussion for a standardized ontology, that improves the exchange of genealogical data and facilitates automatic processing. With such an ontology, standard software tools can be used to encode integrity checks on the data and perform intelligent processing.

References

[1] Family History Department of The Church of Jesus Christ of Latter-day Saints, *The GEDCOM Standard* (Release 5.5), 2 January 1996 (http://www.familysearch.org/GEDCOM/GEDCOM55.exe)

[2] *GENTECH Genealogical Data Model: A Comprehensive Data Model for Genealogical Research and Analysis* (version 1.1), May 29, 2000 (https://www.ngsgenealogy.org/ngsgentech/projects/Gdm/Gdm.htm)

[3] G. Antoniou and F. van Harmelen, *A* Semantic Web *Primer*, London and Cambridge 2004.

[4] http://www.w3.org/RDF/

[5] http://www.w3.org/TR/owl-semantics/

[6] N. Guarino, 'Formal Ontology in Information Systems', in: *Proceedings of FOIS98*, Trento, Italy, 6-8 June 1998. Amsterdam, IOS Press, pp. 3-15.

[7] J.F. Allen, 'Maintaining knowledge about temporal intervals', in: *Communications of the ACM* 26, 832-843

DIMITO: Digitization of rural microtoponyms at the Meertens Instituut

Douwe Zeldenrust

Introduction

The end of 2004 saw the start of a brand new project at the Meertens Institute. Its name was Dimito, short for the DIgitization of rural MIcroTOponyms. Rural microtoponyms is the collective term for the names of small entities in both natural and man-made landscape. The first category covers all sorts of rugged features, such as moors, natural forests and marshes, as well as streams, lakes etcetera. The second covers cultivated landscape and includes individual parcels as well as arable land, grazing land and man-made forests. This collection of rural microtoponyms is the largest onomastic collection at Meertens. Often, the phenomenon is designated by the word 'field name', but this paper will use the word 'microtoponym'.[1]

For thirty years, the Meertens Institute has been gathering data on the plethora of microtoponyms in the Netherlands. This unique material comes mainly on handwritten cards which state the name, the origin of the name, the location and the soil composition and use. The collection contains an estimated 200,000 microtoponyms and over 1,700 topographical maps – mostly from the Kadaster (Dutch Land Registry Office) – upon which the microtoponyms are marked. These maps are referred to as 'field name maps' in the archives of the Meertens Instituut. This term will also be used in this paper.[2]

This collection of microtoponyms is not only an excellent source of information for onomasticians inside and outside the Meertens Instituut, it is also a focus of interest for, amongst others, historians, historical geographers and archaeologists,[3] partly because most of the names relate to parcels of land that have been swallowed up by land consolidation or urban expansion. If the microtoponyms could be digitized with the aid of a geographic information system (GIS) this would facilitate and open up new avenues of research in various disciplines.[4]

Dimito is a pilot project. The key objective is to explore the potential for digitization on the basis of a small sample from the available material. The first part of this paper describes the cards and the field name maps. The second addresses the question of digitization. The third reviews the new opportunities offered by the digital database. The paper ends by answering the question that prompted the pilot in the first place: is it useful and feasible to digitize the entire collection?

Section 1: Cards and field name maps

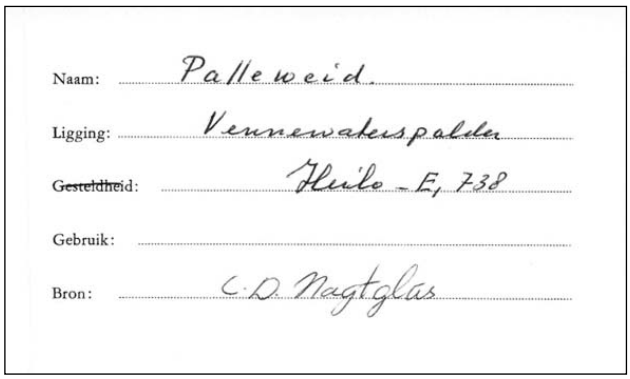
The onomasticians at the Meertens Instituut specialize in the study of proper names. The onomastics discipline consists basically of two subdomains: antroponymy (the study of personal names) and toponymy (the study of place names). Other names belonging to

1 Doreen Gerritzen, *Veldnamen in Noord-Nederland. Een pilot voor een multidisciplinaire database*. Subsidy application for the Digitization Fund (unpublished, 2003). Marc van Oostendorp initiated this project together with Doreen Gerritzen. I am indebted to both of them for their comments on this paper.

2 See the archive of the Meertens Instituut, collection no. 49, collection of field name maps ca. 1860 - 1964 and s.a.

3 H. Beijers (*et al.*), *Veldnamen als historische bron*, een handleiding voor methodisch onderzoek ('s-Hertogenbosch 1991).

4 On 25 April 2003 a workshop was organized at the Meertens Instituut on the study of microtoponyms in the twenty-first century. Presentations were held by, amongst others, Nico Bakker from the Ordnance Survey Office, Jelle Vervloet from Wageningen University and Hans Mol from the Fryske Akademy. See: http://www.meertens.nl/books/veldnamen


Naam: Palleweid
Ligging: Vennewaterspolder
Gesteldheid: Heide - E, 738
Gebruik:
Bron: C.D. Nagtglas

Example 1. A card from the Municipality of Heiloo relating to the microtoponym 'Palleweid'

businesses, organizations, pets etcetera can also feature in the research. This pilot relates to the collection of microtoponyms which Meertens has been building up since 1948.[5]

This collection is the result of the work and commitment of Meertens staff and individuals who donated items of interest. It includes correspondence with researchers, questionnaires, documents about microtoponyms at sites in the Netherlands, rough versions of maps, cards and other letters and documents.[6]

In the course of time one whole collection was compiled from this multiplicity of sources. The data was filtered out and then recorded on around 200,000 cards and inserted on 1,749 accompanying field name maps. There is space on each card for five pieces of data: the name plus (usually) information on the location, the soil composition, the soil use and the origin of the name. The name is marked on the field name map so that the geographical location can be traced.[7]

The cards are filed in drawers and are duplicated. The microtoponyms can be accessed under place name or in alphabetical order per province. There are no place names in the latter system. In principle, the cards in both databases are copies of each other. The pilot made use of the database that can be accessed by place name. This conferred an added benefit: searches could also be performed by place name in the digital version.[8]

Various problems arose when we tried to digitize the cards and the field name maps. To begin with, the 200,000 or so cards had mostly been filled in by hand and were nowhere near as legible as they might have been, so they could not be machine-processed with an OCR program (optical character recognition).[9] The upshot was that the text on the cards had to be entered manually into the database.

Second, the cards had not been systematically filled in. For instance, in the example above, 'composition' has been crossed out and the writer has filled in a reference to one of the field name maps. 'Use' has been left blank. This pattern repeated itself across most of the 400 cards from the Municipality of Heiloo (i.e. the cards used for the pilot). Then, there was the problem of legibility. The people who were recording the entries in the database sometimes had to revert to an educated guess. It was not always clear whether the letters were upper-case or lower-case and punctuation marks had been used on a few cards and left out on the rest. Capital letters and punctuation marks can be important in linguistic research.

Third, the microtoponyms in the sample were undated. They had been collected by individuals or had found their way into the collection from books and archives. So, there was no way of telling whether they were recent or several centuries old. In other words, a microtoponym might just as easily stem from the 16th century as the 20th century.Time stratification is impossible without dates. Researchers will have to take account of this when consulting and comparing the information.

In addition, the microtoponyms could not be exactly located on the basis of, for example, geographical coordinates. This is an essential part of a geographic information system. However, all was not lost, as most of the cards contained a reference to one of the field

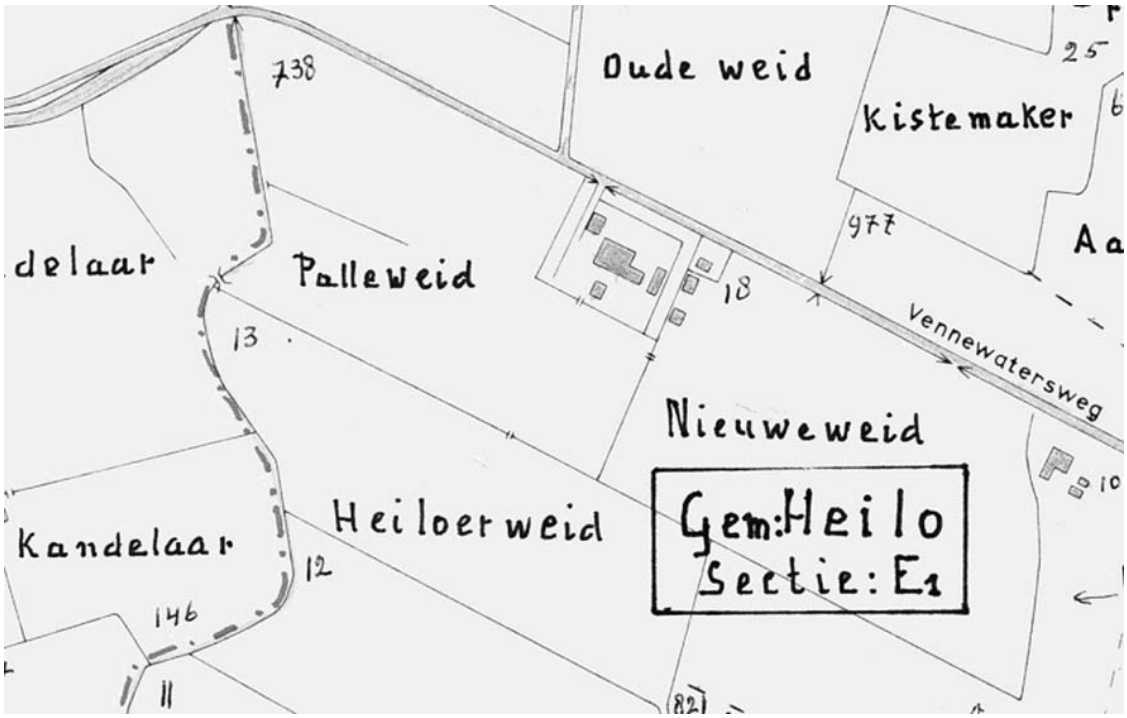
5 Needless to say the collection at the Meertens Instituut does not pretend to be complete. There are sure to be many microtoponyms in the Netherlands which are not in the collection. For more information on the Meertens Instituut see: http://www.meertens.nl

6 See above, the archive of the Meertens Instituut, collection no. 99, field name collection 1941- 1992.

7 Meertens Instituut archive, collection no. 49, field name maps.

8 *Ibid.*

9 OCR converts digital images of printed texts or tables into signs which can be further processed by other computer programs (word-processors, databases, spreadsheets). See also: http://www.niwi.nl/nl/geschiedenis



Example 2. Part of a field name map, belonging to the Palleweid, from the Municipality of Heiloo

name maps, which meant that a position could be determined if the microtoponym was recorded on the map. But neither the size of the field nor its exact location could be ascertained in many cases. Though areas (based on *e.g.* land registry data) were usually marked on the field name maps, they did not always correspond with the microtoponym. It is not impossible for a microtoponym from the 16th century to be marked on a land registry map from the early 20th century. If so, it is most unlikely that the boundaries will still be the same. Finally, there was a distinct possibility that not all the microtoponyms were marked on the field name map and that the number of microtoponyms on the map did not correspond with the information on the cards.

The digitization of the field name maps presented its own problems. The maps had been entered in a database in the past in an attempt to impose some order on the collection.[10] The maps turned out to be highly diverse. Most came from the Land Registry Office, some came from the water authorities and a few had no source reference. Sometimes, dates and scales were missing. Some maps were poor copies, while others turned out to be originals. Some overlapped, and there was absolutely no doubt that they did not cover the whole of the country. Though it was easy enough to digitize the maps, it was a lot more problematic trying to fit them into a geographic information system. Old maps often have imperfections which are copied to the digital environment, and hence, to the location of the microtoponyms. The so-called 'accuracy' of the field name maps turned out to be an illusion.

Section 2: Digitizing the microtoponyms

The problems connected with the digitization of the microtoponyms at the Meertens Instituut also feature in the literature: specifically, determining the exact location of the microtoponym (essential in a geographic information system). Onomastician Rob Rentenaar wrote: 'The more special categories in the sources relevant to toponymy include those created by the onmasticians themselves, such as records, questionnaires and field name maps. In theory, these are the most reliable sources that a researcher could wish for.

10 See the field name map database of Leendert Brouwer from the Meertens Instituut.

After all, the information has been recorded straight from the mouths of the users with the sole aim of obtaining the purest possible toponymical data. All the same, it can do no harm to exercise caution. Memories can change over the years (...).'[11]

This quote from Rentenaar not only suggests that it is not entirely possible to assign an exact position to a microtoponym, it offers a perspective for the problems. Within the context of the quote, the microtoponym collection of the Meertens Instituut can be seen as one of the '(...) categories in the sources relevant to toponymy (...)'.[12] And, as usual, that one unique source has an upside and a downside, so researchers will have to check out its reliability. This is where the digital database can prove useful. The geographic information system in which the microtoponyms will eventually end up will add no extra information to the current collection or make it more accurate. Its only function is to provide better and alternative access to the collections and to facilitate comparison.[13]

With this as the starting point it is possible to digitize the microtoponym collection. Researchers can even check the source by, for example, consulting scans of the originals. They can study the card and (if available) the field name map for each microtoponym. If necessary, the original can always be retrieved, but the quality of the digital copies would be so good that the original could remain undisturbed.

The two components of the microtoponym collection, the cards and the field name maps, would be digitized separately and linked in a geographic information system. The information on the cards and on the field name maps was entered into a database, which would ultimately consist of three groups of data: the data on the cards, followed by the data on the field name maps and, finally, administrative data stating, for example, the creation date of the record.

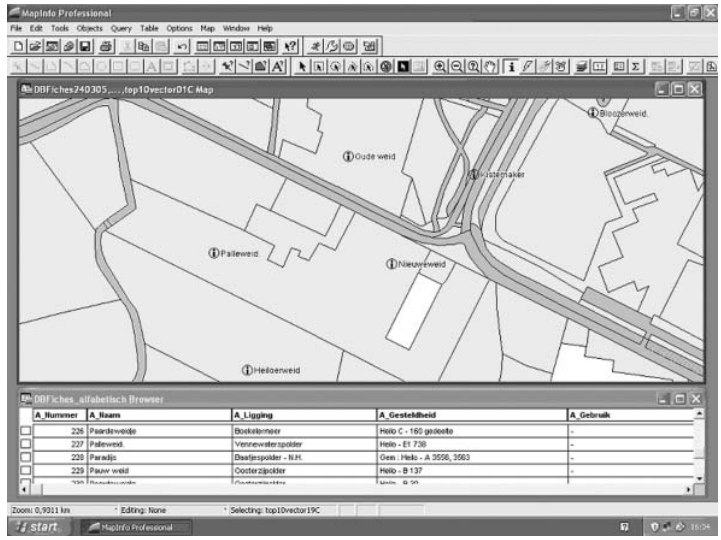
The database had a total of 27 fields. We shall begin by describing the fields for the cards. Each card was assigned a unique number. Obviously, the information on the card would be repeated in the database. The next five fields were: name, location, composition, use and source. The seventh field was for recording any notes on the back of the card. This applied to roughly 8 % of the 400 cards from the Municipality of Heiloo. The eighth field was for the name of the location or the municipality. The ninth was reserved for a computer path for the image of the fiche. The tenth was for the path for the reverse image of the card (if any). This way, an image of the card could be retrieved for each record. The last two fields in the group were reserved for the x and y coordinates of the microtoponym. The search method for these coordinates is explained later.

The next group of database fields were for the field name maps. All eleven fields were copied from the database of the field name maps: identification number, cluster number (for specific dialects in specific regions), cabinet number, specification (often geographic), toponyms (for indicating the presence of toponyms), main card (the database further explanation), title, field name map (the database offers no further explanation), place name, province, and finally, a computer path for the image of the map so that a map could be found for every microtoponym.

Last but not least, the administrative part. This consisted of four fields: one for the entry date of the card, one for the name of the person who entered it, one to indicate whether the microtoponym came only from a card, could only be pinpointed on a map, or had both a card and a position on the map. Finally, field 27 was reserved for remarks. Needless to say, more fields could be added at a later date if required.

As mentioned above, two fields in the database were reserved for an x and a y coordinate. The current collection of microtoponyms does not have geographic coordinates. An accurate system of geographic determination is essential in a GIS, not least for the exchange of geographical information. Otherwise, it is merely an extensive database. [14]

11 Rob Rentenaar, 'Plaatsnamen in historische bronnen', in: *Naamkunde,* no. 2, Vol. 34 2002 (Leuven 2002) 144.
12 *Ibid*
13 Gerritzen, *Veldnamen*
14 For more information see: Ian N. Gregory, *A Place in History,* 'A guide to using GIS in Historical Research' (Oxford 2003) and Paul Ell & Ian N. Gregory, Adding a new dimension to historical research with GIS, in: *History and Computing,* Vol. 13 no. 1 (Edinburgh 2001).



Example 3: The microtoponyms in a geographic information system (MapInfo). The base layer is the TOP 10 Vector Map of the Netherlands; 'i' marks the spot where the microtoponym has been 'pinned'. The database contains the data belonging to the microtoponym. Clicking on 'i' enables you to zoom in on the image of the card of the Palleweid (see Example 1) and the field name map (see Example 2).

Geo-referencing was (and still is) the easiest and most economic technique to assign geographic co-ordinates to the field name maps (and the cards in the next step). Most GIS software has a function that enables this technique to be applied.[15] What happens is that first a map is scanned (in this case a field name map) and the image is projected onto a map of the Netherlands which has x and y coordinates. The pilot used the TOP 10 Vector Map[16]. The image of the field name map is then 'pinned' to the vector map, which gives each point on the field name map a coordinate. It sounds simple but, as mentioned earlier, some of the field name maps have no source references or are poor copies. Old maps can also be seriously flawed. So, positioning can be an awkward and delicate task.

Once the field name maps were geo-referenced, the individual microtoponyms could be assigned an x and a y coordinate. The field had to be manually pinned into the map and the coordinates added at the right record. One extra advantage of this technique is that a microtoponym gets one coordinate. In other words, it is assigned to a point and not to an area. Notwithstanding the inaccuracies in the maps mentioned in paragraph one, Rentenaar suggests that the data is not particularly accurate either. Once a microtoponym has been linked to one point in a geographic information system, there are various ways of adjusting the accuracy; for instance, by marking out a radius of 100 metres within which the microtoponym is valid. It is also possible to compare and exchange files. One click on the mouse will retrieve an image of the card and the field name map for every microtoponym. Hence, researchers can consult the original source in each case.

Section 3: Microtoponyms and openings for research

Having established that it is technically feasible to digitize the microtoponym collection of the Meertens Instituut, this section will discuss the new research openings offered by the database itself and the potential for exchange with other geographic information systems.

15 In this pilot we opted for MapInfo 7.8. First, because MapInfo is already being used at the Meertens Instituut for other projects, so we were already familiar with it, and second, because it met the project criteria. A database could be built which could be adapted to different formats and which also allowed the import of various database formats. The package could also be used for geo-referencing. Last but not least, it was cheaper than the alternative, ArcView.

16 The TOP 10 Vector is the most detailed digital database at the Ordinance Survey Office. The Meertens Instituut is licensed to use it.

The data can now be consulted by other means besides the cards. Qualitative, quantitative and geographical data can be retrieved and displayed. For example, you might want to search for specific microtoponyms, such as all the 'gallows fields', in a specific area. These fields can now be shown on a map of the Netherlands, making the spread visible. Or you might want to look at the geographical distribution of the word *weid* (meadow). There are many instances of *weid* in the Municipality of Heiloo (North Holland) whereas, the province of Brabant might be more inclined to use the world *veldje* (field). You could then try to ascertain where the change occurred and whether it was gradual or abrupt.[17]

Other digital files can also be used as a base layer: a map showing land use could be used instead of the TOP 10 Vector Map. Alterra, part of Wageningen University, has a digital map showing soil use in the Netherlands in 1900.[18] It is now possible, on the basis of the microtoponyms, which are known to sometimes indicate how the soil was used, to investigate the qualitative and quantitative aspects of this relationship.[19]

Moreover, data can be exchanged. The recently launched web site www.kich.nl (**K**nowledge **I**nfrastructure for **C**ultural **H**istory) provides integrated information from four knowledge institutes. The Netherlands Department for Conservation, the *Rijksdienst voor het Oudheidkundig Bodemonderzoek* (government department for archaeological investigation), the Knowledge Directorate of the Ministry of Agriculture, Nature and Food Quality and Alterra (Wageningen University) have unlocked, linked, combined and organized information on cultural history. It is now possible, for example, to select the existing or archaeological monuments in an area and display them on a map.[20] If such a map is used as a base layer for the microtoponym database, the relationships between the microtoponyms and monuments can be explored.

This research serves at least two purposes. First, it enables us to determine how often microtoponyms in a specific area refer to, say, archaeological monuments. Second, it helps us to perform qualitative research. Does the name 'Roman Mound' really indicate Roman remains? And the sword cuts both ways. If there is indeed a demonstrable relationship, it may be that a microtoponyn with an archaeological background points to undiscovered remains and can create some intriguing puzzles for planners, policymakers, designers and other professionals (which KICH names as its main target group).[21] So, onomastic research can extend into other disciplines.

Conclusion

It would, in general, be possible to digitize the microtoponym collection at the Meertens Instituut and set up a geographic information system (GIS), even though some aspects would present a problem. One serious obstacle is the determination of the location of microtoponyms, despite the presence of field maps. A geographic information system requires exact positioning. This is unattainable with a database of field name maps, many of which are decades old and have very little affinity with modern digital maps. Further, the statement by onomastician Rentenaar suggests that human sources of information are not always reliable. We tackled these problems by assigning one point to a microtoponyn instead of an area, and then demarcating a 100-metre radius around it. This leaves space for further adjustments, even for individual cases.

To prevent deterioration in the original material we decided to link each microtoponym to a scanned image of the card containing the data and of the field name map. This means that (good quality) digital copies of the original material can be consulted for each microtoponym. The database can also be accessed from other perspectives, besides the geographic. It contains the data from the cards and other fields that can be filled thanks to the available material.

The digital source that has developed in the meantime does not only provide easier access to the collection, it opens up opportunities for new research questions. Paper data can only be accessed through the cards; this data can be consulted in different ways. Qualitative, quantitative and geographic data can be requested and displayed. One could, for example, search for all the 'gallows fields' in a specific area and show them on a map.

17 For a classification of the field names see: M. Schönveld, *Veldnamen in Nederland (*Amsterdam 1950).
18 See: http://www.hgnnederland.nl
19 Schönveld, bestand 73-87.
20 Zie: www.kich.nl
21 Ibidem.

It also opens new opportunities for interdisciplinary research. Onomastics can create new insight in tandem with other disciplines. Other maps could be used as a base layer for the field names; say, a map of the archaeological monuments in the Netherlands – which would shed light on the relationship between the microtoponyms and monuments. If this relationship were demonstrated, then some microtoponyms might even lead the way to undiscovered archaeological remains. This would provide a valid reason for exploring an area further or conserving it.

Finally, the last issue of *Naamkunde* featured an article in which Hans Bennis, Director of the Meertens Instituut, expressed concern about the future of onomastics as a discipline. He ended with an appeal: ' (...) Fellow onomasticians, draw attention to the importance and the role of onomastics in the academic scheme of things!'[22] Perhaps, a digital database of microtoponyms offering new research opportunities will go some way towards communicating that message, and win onomastics a place alongside other geographically-oriented humanities.

References

Archief Meertens Instituut, collection no. 49, field name map collection ca. 1860 - 1964 and s.a.

Archief Meertens Instituut, collection no. 99, field name collection 1941 - 1992.

Meertens Instituut, field name map database, Leendert Brouwer.

http://www.hgnnederland.nl

http://www.kich.nl

http://www.meertens.nl

http://meertens.nl/books/veldnamen

http://www.niwi.nl/nl/geschiedenis

Beijers, H. (*et al.*), *Veldnamen als historische bron, een handleiding voor methodisch onderzoek* ('s-Hertogenbosch 1991).

Bennis, Hans, 'Naamkunde als discipline', in: *Naamkunde,* No. 2, 34ste jaargang 2002 (Leuven 2002).

Ell, Paul & Gregory, Ian N., 'Adding a new dimension to historical research with GIS', in: *History and Computing,* Vol. 13, no. 1 (Edinburgh 2001).

Gerritzen, Doreen, V*eldnamen in Noord-Nederland.* Een pilot voor een multidisciplinaire database. Subsidie aanvraag Digitaliseringsfonds (unpublished, 2003).

Gregory, Ian N., *A place in History, A guide to using GIS in historical research* (Oxford 2003).

Rentenaar, Rob, 'Plaatsnamen in historische bronnen', in: *Naamkunde,* No. 2, Vol 34 2002 (Leuven 2002) pp. 137-148.

22 Hans Bennis, 'Naamkunde als discipline', in: *Naamkunde,* No. 2, Vol 34 2002 (Leuven 2002) 135.